LNCS 6979

Giuseppe Maino
Gian Luca Foresti (Eds.)

Image Analysis and Processing – ICIAP 2011

16th International Conference
Ravenna, Italy, September 2011
Proceedings, Part II

2 Part II

IAPR

Springer

Lecture Notes in Computer Science 6979

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

Giuseppe Maino   Gian Luca Foresti (Eds.)

Image Analysis and Processing – ICIAP 2011

16th International Conference
Ravenna, Italy, September 14-16, 2011
Proceedings, Part II

Springer

Volume Editors

Giuseppe Maino
Università di Bologna
Facoltà di Conservazione dei Beni Culturali
Via Mariani 5, 48100 Ravenna, Italy
E-mail: giuseppe.maino@unibo.it

Gian Luca Foresti
Università di Udine
Dipartimento di Matematica e Informatica
via delle Scienze 206, 33100 Udine, Italy
E-mail: gianluca.foresti@uniud.it

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-24087-4 e-ISBN 978-3-642-24088-1
DOI 10.1007/978-3-642-24088-1
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011936230

CR Subject Classification (1998): I.4, I.5, I.3.5, I.2.10, I.2.6, H.3, F.2.2

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

This volume collects the papers accepted for presentation at the International Conference on Image Analysis and Processing (ICIAP 2011), held in Ravenna, Italy, September 14–16, 2011. ICIAP 2011 was the 16th event in a series of conferences organized biennially by the Italian Member Society of the International Association for Pattern Recognition (IAPR). The aim of these conferences is to bring together international researchers for the presentation and discussion of the most recent advances in the fields of pattern recognition, image analysis, and image processing. Following the successful 2009 conference in Vietri sul Mare, ICIAP 2011 was held in the magnificent city of Ravenna, an historical city famous for its artistic and cultural heritage. The 16th ICIAP conference was organized jointly by the Faculty of Preservation of Cultural Heritage of the University of Bologna and the Department of Mathematics and Computer Science (DIMI) of the University of Udine.

Topics for ICIAP 2011 included Image Analysis and Processing, Pattern Recognition and Vision, Multimodal Interaction and Multimedia Processing, Cultural Heritage, and Applications.

There were 175 submissions. Each submission was reviewed by two Program Committee members. The committee decided to accept 121 papers, divided into 10 oral sessions (44 papers) and three poster sessions (77 papers).

The program included a special session on “Low Level Color Image Processing” (organized by M. Emre Celebi, Bogdan Smolka, Gerald Schaefer, and Raimondo Schettini), a demo session, and four invited talks by Jake K. Aggarwal (University of Texas, Department of Electrical and Computer Engineering, USA) on *Recognition of Human Activities*, Horst Bunke (University of Bern, Institute of Computer Science and Applied Mathematics, Switzerland) on *Bridging the Gap between Structural and Statistical Pattern Recognition*, Roberto Cipolla (University of Cambridge, Department of Engineering, UK), on *Novel Applications of 3D Shape from Uncalibrated Images*, and Kevin Karplus (University of California, Santa Cruz, Department of Biomolecular Engineering, USA) on *Protein Structure and Genome Assembly Tools.* These lectures survey established approaches, recent results and directions of future works of different topics of recognition of human activities, structural and statistical pattern recognition, computational vision, bioinformatics, and biomolecular engineering.

Three tutorials were offered, on “Image and Video Descriptors” (by Abdenour Hadid), on “Beyond Features: Similarity-Based Pattern Analysis and Recognition” (by Edwin R. Hancock, Vittorio Murino, and Marcello Pelillo), and on “Video Analytics on Reactive Camera Networks” (by Christian Micheloni).

ICIAP 2011 will also host the First International Workshop on Pattern Recognition in Proteomics, Structural Biology and Bioinformatics, PR PS BB 2011, organized by Virginio Cantoni and Giuseppe Maino.

During the conference, the Caianiello Prize, in memory of Prof. E. Caianiello, was awarded to the best paper by a young author, as at previous events. Also, a prize was awarded to the best paper presented to the conference.

We wish to thank the Italian group of researchers affiliated to the International Association for Pattern Recognition (GIRPR) for giving us the opportunity to organize this conference. We also thank the International Association for Pattern Recognition for the endorsement of ICIAP 2011. A special word of thanks goes to the Program Chairs, to the members of the Program Committee and to the reviewers, who contributed with their work to ensuring the high-quality standard of the papers accepted to ICIAP 2011.

Special thanks go to Claudio Piciarelli, who made a fundamental contribution to this conference, helping in managing, working on, and resolving those many problems that a large event like this presents.

Local organization for events and accommodation was managed by Carla Rossi of the Fondazione Flaminia and Daniela Raule of the NEREA-AIDA spin-off. We are indebted to the Fondazione Flaminia for financial and organization support. A special thanks goes to the members of the Local Organizing Committee, Roberta Menghi and Mariapaola Monti, who also took care of the graphic aspects of the event, Elena Nencini, Lorenza Roversi, and Lisa Volpe for their indispensable contribution to the organization and their help and availability to solve the many practical problems arising during the preparation of ICIAP 2011. Finally, Sara Armaroli, Donatella Lombardo, Mariapaola Monti, and Liu Wan are the young artists that have lent themselves to realize the Vision&Art exhibition accompanying ICIAP 2011.

September 2011

Giuseppe Maino
Gian Luca Foresti

Organization

Organizing Institutions

Alma Mater Studiorum, Università di Bologna
Università degli Studi di Udine

General Chairs

Giuseppe Maino	ENEA and University of Bologna, Italy
Gian Luca Foresti	University of Udine, Italy

Program Chairs

Sebastiano Battiato	University of Catania, Italy (Image Analysis and Processing)
Donatella Biagi Maino	University of Bologna, Italy (Cultural Heritage and Applications)
Christian Micheloni	University of Udine, Italy (Pattern Recognition and Vision)
Lauro Snidaro	University of Udine, Italy (Machine Learning and Multimedia)

Publicity Chair

Claudio Piciarelli	University of Udine, Italy

Steering Committee

Virginio Cantoni, Italy
Luigi Cordella, Italy
Alberto Del Bimbo, Italy
Marco Ferretti, Italy
Fabio Roli, Italy
Gabriella Sanniti di Baja, Italy

Program Committee

Jake K. Aggarwal, USA
Maria Grazia Albanesi, Italy
Hlder J. Araújo, Portugal
Edoardo Ardizzone, Italy
Prabir Bhattacharya, USA
Alessandro Bevilacqua, Italy
Giuseppe Boccignone, Italy
Gunilla Borgefors, Sweden
Alfred Bruckstein, Israel
Paola Campadelli, Italy
Elisabetta Canetta, UK
Andrea Cavallaro, UK
Rémy Chapoulie, France
M. Emre Celebi, USA
Rita Cucchiara, Italy
Leila De Floriani, Italy
Claudio De Stefano, Italy
Pierre Drap, France
Jean Luc Dugelay, France
Ana Fred, Portugal
Maria Frucci, Italy
André Gagalowicz, France
Giorgio Giacinto, Italy
Edwin Hancock, UK
Francisco H. Imai, USA
Rangachar Kasturi, USA
Walter Kropatsch, Austria
Josep Lladòs, Spain
Brian C. Lovell, Australia
Rastislav Lukac, Canada
Angelo Marcelli, Italy
Simone Marinai, Italy
Stefano Messelodi, Italy
Vittorio Murino, Italy
Mike Nachtegael, Belgium
Michele Nappi, Italy
Hirobumi Nishida, Japan
Jean-Marc Ogier, France
Marcello Pelillo, Italy
Alfredo Petrosino, Italy
Maria Petrou, Greece
Matti Pietikäinen, Finland
Giuseppe Pirlo, Italy
Fabio Remondino, Switzerland
Hanan Samet, USA
Carlo Sansone, Italy
Silvio Savarese, USA
Gerard Schaefer, UK
Raimondo Schettini, Italy
Linda Shapiro, USA
Filippo Stanco, Italy
Massimo Tistarelli, Italy
Alain Trémeau, France
Roberto Tronci, Italy
Adrian Ulges, Germany
Cesare Valenti, Italy
Mario Vento, Italy
Daniele Visparelli, Italy
Domenico Vitulano, Italy
Yehezkel Yeshurun, Israel
Marcel Worring, The Netherlands
Lei Zhang, Hong Kong, China
Primo Zingaretti, Italy
Galina I. Zmievskaya, Russia

Additional Reviewers

Lamberto Ballan
Silvia Bussi
Elena Casiraghi
Paul Ian Chippendale
Luca Didaci
Giovanni Maria Farinella
Francesco Fontanella
Alessandro Gherardi
Cris Luengo Hendriks
Michela Lecca
Paola Magillo
Iacopo Masi
Carla Maria Modena
Daniele Muntoni
Gabriele Murgia
Paolo Napoletano

Francesca Odone
Federico Pernici
Maurizio Pili
Giovanni Puglisi
Ajita Rattani
Elisa Ricci
Reza Sabzevari
Riccardo Satta
Giuseppe Serra
Nicola Sirena
Lennart Svensson
Francesco Tortorella
Ingrid Visentini
Erik Wernersson
Matteo Zanotto

Local Organizing Committee

Roberta Menghi
Mariapaola Monti
Carla Rossi
Lorenza Roversi
Lisa Volpe
Basilio Limuti

Endorsing Institutions

Italian Member Society of the International Association for Pattern Recognition – GIRPR
International Association for Pattern Recognition – IAPR

Sponsoring Institutions

Fondazione Flaminia, Ravenna
Ordine della Casa Matha, Ravenna

Table of Contents – Part II

Image and Video Analysis and Processing

Applications

Table of Contents – Part I

Image Analysis and Representation

Image Segmentation

Pattern Analysis and Classification

Forensics, Security and Document Analysis

Video Analysis and Processing

Biometry

Shape Analysis

Low-Level Color Image Processing

Applications

Medical Imaging

Image Analysis and Pattern Recognition

A Visual Blindspot Monitoring System for Safe Lane Changes

Jamal Saboune[1], Mehdi Arezoomand[1], Luc Martel[2], and Robert Laganiere[1]

[1] VIVA Lab, School of Information Technology and Engineering,
University of Ottawa, Ottawa, Ontario, K1N 6N5, Canada
[2] Cognivue Corporation, Gatineau, Quebec J8X 4B5 Canada
{jsaboune,marezoom,laganier}@site.uottawa.ca, lmartel@cognivue.com

Abstract. The goal of this work is to propose a solution to improve a driver's safety while changing lanes on the highway. In fact, if the driver is not aware of the presence of a vehicle in his blindspot a crash can occur. In this article we propose a method to monitor the blindspot zone using video feeds and warn the driver of any dangerous situation. In order to fit in a real time embedded car safety system, we avoid using any complex techniques such as classification and learning. The blindspot monitoring algorithm we expose here is based on a features tracking approach by optical flow calculation. The features to track are chosen essentially given their motion patterns that must match those of a moving vehicle and are filtered in order to overcome the presence of noise. We can then take a decision on a car presence in the blindspot given the tracked features density. To illustrate our approach we present some results using video feeds captured on the highway.

1 Introduction

Car accidents on the highways are a big factor of mortality and can cause severe injuries. Actually, drivers nowadays are getting more concerned about safety features in their cars and thus are willing to pay the cost of acquiring safer vehicles. On the other hand, the public services are interested in reducing the mortality rate on the roads considered nowadays as an indicator of the quality of life. Motivated by economic factors, a new research domain has thus emerged in the recent years; It is known as pre-crash sensing. The research in this domain, conducted by car manufacturers as well as by public research institutions, aims to make the vehicles safer and as a result reduce the number of crashes and their severity. The main threat for a driver on the highway comes from the surrounding cars especially when he is not aware of their close presence. In fact one of the the main features of an onboard car safety system is to detect the presence of a close car in the driver's blindspot (Figure 1) and warn the latter about it. This information can help the driver in a lane change situation and affect his decision to perform this task. In this paper we present a simple and fast approach for blindspot monitoring using computer vision. This feature is essential and can help preventing many risky driving situations.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 1–10, 2011.

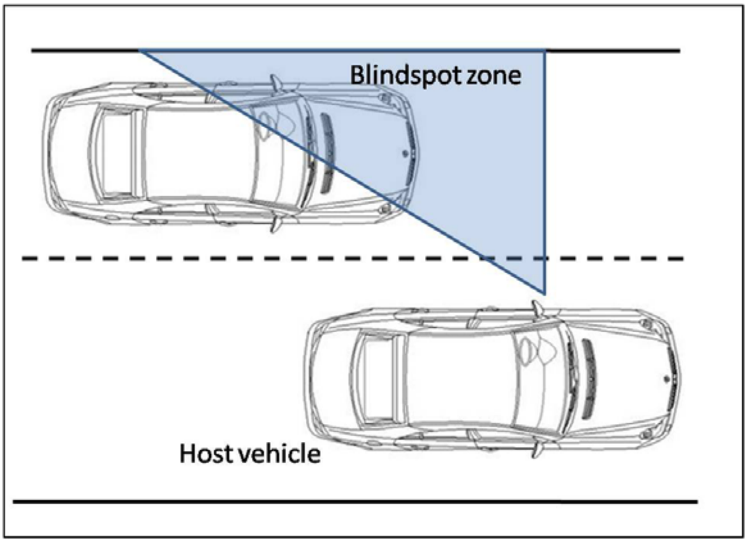


Fig. 1. The blindspot zone description: We define the blindspot of a driver as the zone he can not see through his side and rear view mirrors

The blindspot monitoring problem is a problem of a car detection in a given zone surrounding the host car. This car detection task which is the initial step to accomplish in any collision avoidance system, has been widely adressed. The first generation of collision avoidance systems is based on using a radar technology. These systems adopt diverse technologies such as infrared, ultrasonic waves [23], sonars [15] or laser scanners [27] in order to detect the presence of any object in the range of the sensors embedded in the car's body. Some of the radars used can also detect the shape of the object. However the radars used have a small range and as a result are not able to detect some approaching vehicles. On the other hand, their field of view is reduced and present blindspots. Thus, in order to cover the wide area surrounding the car many sensors are needed which increases the cost of the system. With the recent advances in nanotechnology and in imagery sensing devices the new generation of embedded security system is relying on the use of small cameras installed in different locations of the vehicle; the cameras can have a large field of view which enables the system to detect vehicles moving in a large area and overcome the radars disadvantages. The cameras are also low cost and can be combined with radars to improve the car detection [14,15,16,19,27].

In order to detect a vehicle in the camera feeds, three approaches were adopted in previous works. The first one known as 'knowledge based' relies on recognizing the vehicle in a certain single image given some distinctive features. In fact, vehicles have some distinctive visual features (color, shape etc.) and thus vehicle detection in an image can be reduced to a classical pattern recognition problem. This problem can be solved using a features vector classification technique given a database of learned features vectors representing vehicles and roads. The features used for classification can be of different types. Tsai et al. [25] use color and edge information and a bayesian classifier to resolve this problem. Haar like [10,20,27] and Histogram of Oriented Gradient (HOG) [2,17,20,1] features were also widely used. These distinctive features can then be classified using Support Vector Machine (SVM) [2,17,1] or Adaboost classifiers [10,20,27]. However, vehicles are of different shapes and colors and can be viewd from different angles and under different illumination conditions in videos. Thus, the database should be very wide in order to be inclusive and to have a good recognition. The classification step is also time consuming. Given that, we can say that these algorithms are complex and not well adapted to an on-board system. In order to avoid the learning and classification steps, other methods using a car distinctive features were proposed. Detecting the shadow underneath a car is a sign of a car's presence [8,15,28]. Unfortunately it is not always possible to detect this shadow especially for cars moving far from the camera or in a cloudy or dark environment. This feature can thus be combined with other features such as vertical edges and symmetry rate [15], left and right car borders [8] or lights detection [28] for night situations. Despite those improvements, the vehicles recognition in this type of approaches is not accurate and is highly perturbated by shadows of the background objects (guard rails, trees etc.). Collado et al. [9] constructed geometric models of the car with energy functions including shape and symmetry. This approach succeded in detecting far preceding cars but presented some weakness in detecting lateral and close cars. Wu et al. [29] succeded in detecting cars in the blindspot by comparing the grayscale histogram of the road surface to that of a patch covering the neighbouring lane. This idea is effective in detecting any object whose colour is defferent from that of the road but does not imply that the object is a car.

The second approach known as 'motion based' uses an object's motion information estimated through successive images to detect the vehicle's presence. This idea is motivated by the fact that a vehicle moves relatively to the background with a standard motion pattern. In order to estimate the motion of an object we need to find the correspondances between the features describing it in the successive images. To accomplish this, color, edge and contour information [5,21,24] can be efficient as well as SURF [6] or SIFT [13] features. Spatiotemporal wavelet transforms [26], image entropy [7] and optical flow algorithms [3,11] were also employed for motion estimation. These techniques proved to be less complex and more efficient than the 'knowledge based' ones although they present some errors in specific cases. The motion and knowledge based approaches can also be combined. The third technique is a stereo vision method [4,12] that calculates the disparity map and accomplishes a 3D reconstruction but is very complex and highly inaccurate.

In order to simplify the problem and to develop a fast algorithm that can be easily implemented we decided to adopt a 'motion based' strategy for car detection and tracking in the blindspot. Our method uses an optical flow technique to estimate the motion of some features that would represent a car. These features are well chosen in a way to avoid false detections and thus reduce the algorithm complexity. The developped method will be exposed in section 2. Results and discussion will later be exposed in section 3.

2 Blindspot Car Detection and Tracking

The easiest and most efficient solution to detect a car moving in front of the camera can be accomplished by detecting the zone in the image representing it and then track it using a template matching technique. To detect a vehicle presence in the blindspot, we cannot use the same logic. In fact, in order to accomplish this task efficiently we need to install the camera in a way to be able to see a car when it is aproaching from behind and then passing by the driver. We cannot therefore install the camera on the backside of the car but it should be installed in front. In this scene's configuration, a car shape and contour seen from the front, change continouesly when its approaching or passing by the driver. As a result, a template matching technique would act poorly. In order to overcome this problem we decided to adopt a 'motion based' approach using the optical flow calculation. Our idea is motivated by the fact that when the camera is moving forward, the background objects and the slower cars the driver is passing by, move backward relatively to the camera. On the other hand, the approaching cars which present a threat move with a motion pattern close to that of the camera or faster and thus move relatively forward. (Figure 2).

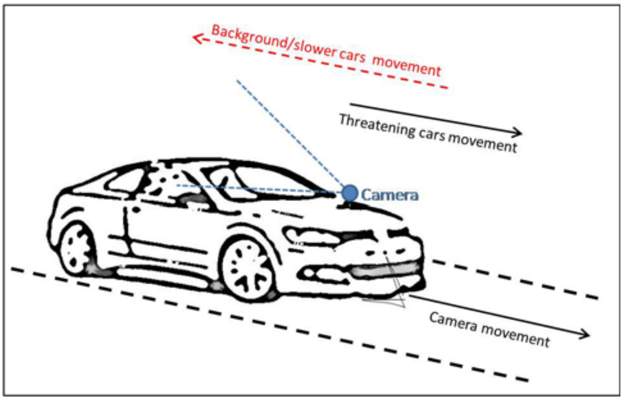


Fig. 2. Configuration of the camera for the blindspot detection problem: The background objects and slower cars move relatively backward and the threatening cars move forward

To estimate the motion of the moving objects in the successive images we use an optical flow calculation approach on a group of features describing these objects. We opted to use the Shi & Tomasi Good features to track [22] that proved to be efficient for object tracking and to use the pyramidal Lucas - Kanade tracker [18] for optical flow calculation. We then apply a number of filters in order to make sure we are tracking only features representing a vehicle and not noise The different steps of our algorithm are the following:

- We first specify a zone in the grayscale image covering the lane next to the camera; This will be the zone we want to track a vehicle in (the blindspot zone).
- Every 8 frames we extract the features to track in this zone, as we observed that new objects do not appear completely in less than that temporal distance. Thus we are able to reduce the complexity of the method. The extracted features are added to the set of features to track S (resulting from the previous frames).
- By calculating the optical flow of all features in S we then estimate the motion vector of each. We consider that the motion vectors verifying some conditions represent a potential car and we label them as 'valid features'. Else, they are rejected from S and we stop tracking them. These conditions are based on the motion patterns described earlier; In fact if a motion vector forms an angle with the horizontal greater than 30 deg. and smaller than 60 deg. and has a value bigger than 1 pixel it is considered as describing a 'valid feature'. This choice is justified by the fact that objects moving with such motion pattern would represent objects (cars) moving in a similar way to the host car and thus are considered as dangerous. Else if a car in the blindspot zone moves with a motion vector which angle to the horizontal is between -30 and 30 deg or 60 and 120 deg its driver is most probably trying to change lanes to the left (getting far from the host car) or to the right (going behind the host car in the same lane). In all the other configurations of this angle the object is either a part of the background or moving slower than the host car and as a result its non-threatening.
- We observe the motion vectors of a 'valid feature' calculated by optical flow over three frames. If these vectors respect the motion conditions we established earlier, for the three frames, we label the corresponding feature as a 'potential feature' and we keep it in our set S. If not, it would be rejected as well and its tracking is stopped. We thus make sure that we eliminate objects having an inconsistent movement.
- In order to avoid tracking features representing some noise we impose an additional condition on the tracked 'potential features'. If a car is present in a certain zone of the blindspot area, its corresponding features should be as well. As a result, we would have a minimum number of 'potential features' in this zone. On the other hand, if a feature is isolated in the zone it is most likely that it represents noise. To illustrate this idea we divide the blindspot zone in five zones of different sizes and we impose a treshold for each of them (Figure 3). If the number of 'potential features' present in one of the zones is less than the treshold fixed for the zone, these features are rejected. The zone containing the biggest number of features is considered as the one containing the vehicle.
- Despite these strict conditions some noise features were able to survive and we had to add a last condition. It is based on the comparison of the pixels

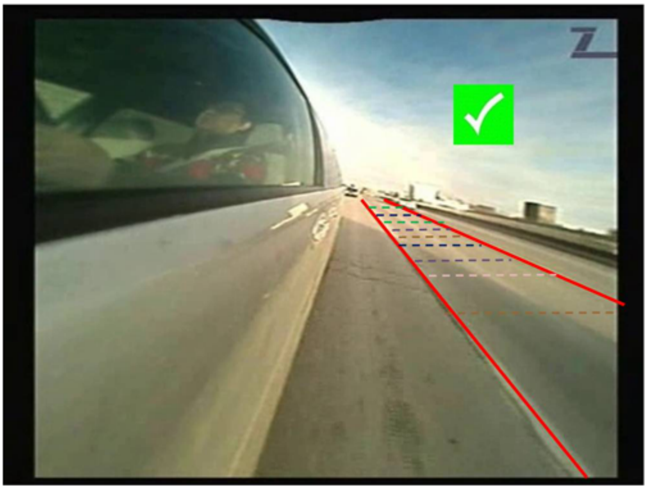


Fig. 3. Image decomposition into zones: The region of the image delimited by the two red lines is considered as the blindspot zone. The dashed colored lines delimit the zones considered as potentially containing a vehicle. Their sizes are estimated given the shape of the vehicle in the next lane and its distance from the camera.

intensity distribution in each zone. When a zone represents the road, the standard deviation of its pixels intensities is small. In opposition, when a car covers the zone, this standard deviation is big. By imposing a treshold on the standard deviation of intensities we were able to eliminate the false positive detections. The features surviving all the conditions are finally labeled as 'vehicle features' and are kept in the set S of features to track.

This scheme proved to be efficient for detecting and tracking all cars moving faster than the camera. In some particular cases where a car tried to pass the camera car but stayed in its blindspot, our system failed. This is due to the fact that in this case the threatening car was moving relatively backward or at the same speed. To solve this issue we decided to keep tracking all the features labeled as 'vehicle features' even if their motion does not respect the conditions established before. In fact we only stop tracking those features in the case they disappear from the image or move with a strict horizontal movement that implies the car is changing lanes.

3 Application and Results

This algorithm was applied to video feeds captured using a commercial CMOS camera fixed on the side mirror of a car moving on a highway. The algorithm was implimented in C++ using the OpenCV library and an Intel Core2 quad processor. The features extraction task, done every 8 frames, took 100 ms to be accomplished. The optical flow calculation and all the filtering steps took

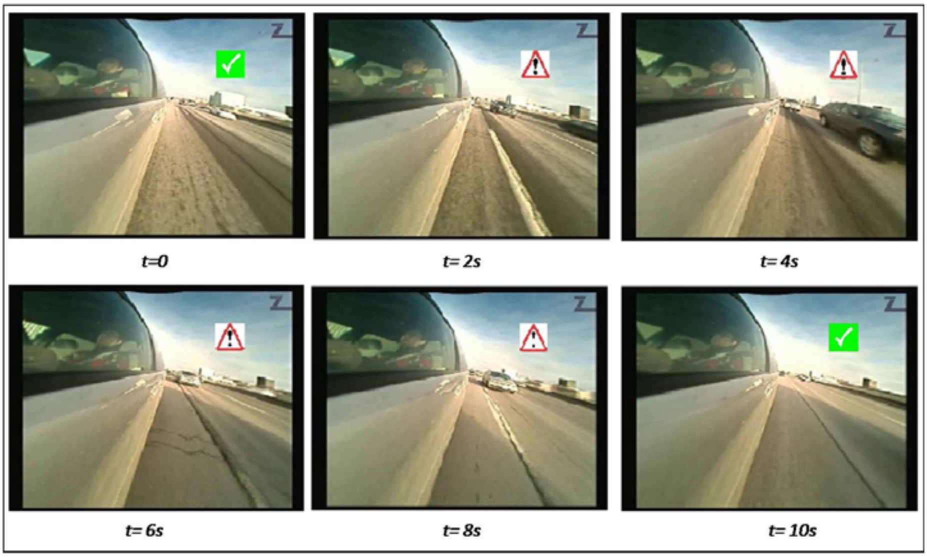


Fig. 4. Blindspot presence detection and tracking result: The images show an exemple of risky situation detection; At $t = 0$ when no car is present in the lane next to the driver no alarm is emitted. As soon as a car enters the zone ($t = 2s$) the system was able to detect it. Another car enters the zone ($t = 4s$) and stays in until the first one disappears ($t = 6s$) The alarm in that case was still valid. We only declare the zone safe (no alram) when the two cars disappear from the lane ($t = 10s$)

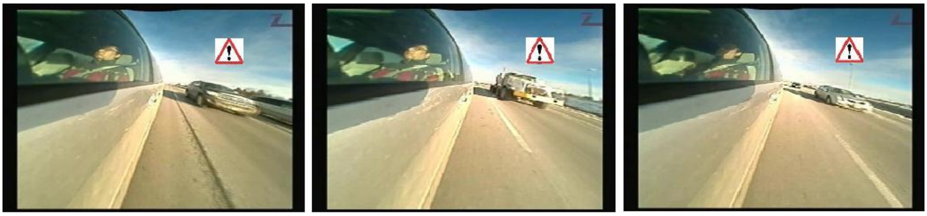


Fig. 5. Blindspot monitoring: The algorithm is efficient in detecting cars of different types and sizes (SUV, truck, standard etc.)

20 ms of calculation time/frame. For 10 000 frames captured, 41 situations of a car presence in the blindspot were encountered. For these 41 situations we were able to detect and track cars in 38 situations without any false positive result. The first situation we missed is caused by the fact that the threatening car passed under a bridge so it was covered by shadow and its tracking was lost momentarily but was detected afterwards as a new threat. The other two missed cars are cars the driver tried to pass by without success. For this particular situation, the cars had a permanent relative backwards movement and we have

to find a new approach for this particular case. Overall the results are very satisfactory (Figure 4) and proved that we are able to detect a risky situation fast and with a simple algorithm. Our approach can also be qualified as generic as our system was able to detect different car types (Figure 5) in opposition to a classification approach where samples of each type have to be included in the learning database.

4 Conclusion

In this paper we presented a new system for blindspot monitoring, intended to be implemented in a car's safety system. Our challenge was therefore to use a simple and fast algorithm but efficient at the same time. We managed to avoid a complex learning - classification technique and came up with a generic solution to detect any type of cars without the need of any database. A car presence in the blindspot was detected based on its features motion patterns. We actually applied a features tracking using optical flow calculation and added some filtering steps to make sure we do not have any false positive detection. By applying these filters we also managed to reduce the number of features to track and as a result the calculation time. The first results were encouraging but we still have to find a solution for a rare particular case. Complexity wise, the algorithm we presented here is simple and fast and can be easily tuned and adapted which makes our system a good candidate to be ported on any microchip as a part of a real time automotive safety system.

References

1. Alvarez, S., Sotelo, M.A., Ocana, M., Llorca, D.F., Parra, I.: Vision-based target detection in road environments. In: WSEAS VIS 2008 (2008)
2. Balcones, D., Llorca, D., Sotelo, M., Gavilin, M., lvarez, S., Parra, I., Ocaa, M.: Real-time vision-based vehicle detection for rear-end collision mitigation systems. In: vol. 5717, pp. 320–325 (2009)
3. Batavia, P., Pomerleau, D., Thorpe, C.: Overtaking vehicle detection using implicit optical flow. In: IEEE Conference on Intelligent Transportation System, ITSC 1997, pp. 729–734 (November 1997)
4. Bertozzi, M., Broggi, A., Fascioli, A., Nichele, S.: Stereo vision-based vehicle detection. In: Proceedings of the IEEE on Intelligent Vehicles Symposium IV 2000, pp. 39–44 (2000)
5. Betke, M., Haritaoglu, E., Davis, L.S.: Real-time multiple vehicle detection and tracking from a moving vehicle. Machine Vision and Applications 12, 69–83 (2000), http://dx.doi.org/10.1007/s001380050126, doi:10.1007/s001380050126
6. Chang, W.C., Hsu, K.J.: Vision-based side vehicle detection from a moving vehicle. In: International Conference on System Science and Engineering (ICSSE), 2010, pp. 553–558 (2010)
7. Chen, C., Chen, Y.: Real-time approaching vehicle detection in blind-spot area. In: 12th International IEEE Conference on Intelligent Transportation Systems, ITSC 2009, pp. 1–6 (2009)

8. Chern, M.Y.: Development of a vehicle vision system for vehicle/lane detection on highway. In: 18th IPPR Conf. on Computer Vision, Graphics and Image Processing, pp. 803–810 (2005)
9. Collado, J., Hilario, C., de la Escalera, A., Armingol, J.: Model based vehicle detection for intelligent vehicles. In: 2004 IEEE Intelligent Vehicles Symposium, pp. 572–577 (2004)
10. Cui, J., Liu, F., Li, Z., Jia, Z.: Vehicle localisation using a single camera. In: 2010 IEEE Intelligent Vehicles Symposium (IV), pp. 871–876 (2010)
11. Diaz Alonso, J., Ros Vidal, E., Rotter, A., Muhlenberg, M.: Lane-change decision aid system based on motion-driven vehicle tracking. IEEE Transactions on Vehicular Technology 57(5), 2736–2746 (2008)
12. Franke, U., Joos, A.: Real-time stereo vision for urban traffic scene understanding. In: Proceedings of the IEEE on Intelligent Vehicles Symposium IV 2000, pp. 273–278 (2000)
13. Jeong, S., Ban, S.W., Lee, M.: Autonomous detector using saliency map model and modified mean-shift tracking for a blind spot monitor in a car. In: Seventh International Conference on Machine Learning and Applications, ICMLA 2008, pp. 253–258 (2008)
14. Kato, T., Ninomiya, Y., Masaki, I.: An obstacle detection method by fusion of radar and motion stereo. IEEE Transactions on Intelligent Transportation Systems 3(3), 182–188 (2002)
15. Kim, S., Oh, S.Y., Kang, J., Ryu, Y., Kim, K., Park, S.C., Park, K.: Front and rear vehicle detection and tracking in the day and night times using vision and sonar sensor fusion. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005), pp. 2173–2178 (2005)
16. Labayrade, R., Royere, C., Gruyer, D., Aubert, D.: Cooperative fusion for multi-obstacles detection with use of stereovision and laser scanner. Autonomous Robots 19, 117–140 (2005), http://dx.doi.org/10.1007/s10514-005-0611-7, doi:10.1007/s10514-005-0611-7
17. Llorca, D.F., Snchez, S., Ocaa, M., Sotelo, M.A.: Vision-based traffic data collection sensor for automotive applications. Sensors 10(1), 860–875 (2010), http://www.mdpi.com/1424-8220/10/1/860/
18. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision, pp. 674–679 (1981)
19. Mar, J., Lin, H.T.: The car-following and lane-changing collision prevention system based on the cascaded fuzzy inference system. IEEE Transactions on Vehicular Technology 54(3), 910–924 (2005)
20. Negri, P., Clady, X., Prevost, L.: Benchmarking haar and histograms of oriented gradients features applied to vehicle detection. ICINCO-RA (1), 359–364 (2007)
21. She, K., Bebis, G., Gu, H., Miller, R.: Vehicle tracking using on-line fusion of color and shape features. In: Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems, pp. 731–736 (2004)
22. Shi, J., Tomasi, C.: Good features to track. In: 1994 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 1994), pp. 593–600 (1994)
23. Song, K.T., Chen, C.H., Huang, C.H.C.: Design and experimental study of an ultrasonic sensor system for lateral collision avoidance at low speeds. In: 2004 IEEE Intelligent Vehicles Symposium, pp. 647–652 (2004)
24. Techmer, A.: Real time motion analysis for monitoring the rear and lateral road. In: 2004 IEEE Intelligent Vehicles Symposium, pp. 704–709 (2004)

25. Tsai, L.W., Hsieh, J.W., Fan, K.C.: Vehicle detection using normalized color and edge map. In: IEEE International Conference on Image Processing, ICIP 2005, vol. 2, pp. II- 598–II-601 (2005)
26. Wang, Y.K., Chen, S.H.: A robust vehicle detection approach. In: IEEE Conference on Advanced Video and Signal Based Surveillance, AVSS 2005, pp. 117–122 (2005)
27. Wender, S., Dietmayer, K.: 3d vehicle detection using a laser scanner and a video camera. Intelligent Transport Systems, IET 2(2), 105–112 (2008)
28. Wu, B.F., Chen, C.J., Li, Y.F., Yang, C.Y., Chien, H.C., Chang, C.W.: An embedded all-time blind spot warning system. In: Zeng, Z., Wang, J. (eds.) Advances in Neural Network Research and Applications. LNCS, vol. 67, pp. 679–685. Springer, Heidelberg (2010)
29. Wu, B.F., Chen, W.H., Chang, C.W., Chen, C.J., Chung, M.W.: A new vehicle detection with distance estimation for lane change warning systems. In: 2007 IEEE Intelligent Vehicles Symposium, pp. 698–703 (2007)

Extracting Noise Elements while Preserving Edges in Spatial Domain

Jalil Bushra, Fauvet Eric, and Laligant Olivier

Le2i Laboratory, Universit de Bourgogne, 12 Rue de la Fonderie, Le Creusot, France
bushra.jalil@u-bourgogne.fr

Abstract. In this paper, we are interested in preserving the sharp transitions and edges present inside the image. Image denoising by means of wavelet transforms has been an active research topic for many years. In this work, we used Lipschitz exponents based on wavelet transform to performs edge preservation by identifying these transitions. The smoothing part was performed by using some heuristic approach utilizing data samples and smoothness criteria in spatial domain with out prior modeling of either the image or noise statistics. The method tries to find the the best compromise between the data and the smoothing criteria based on the type of the transition present. The method has been compared with the shrinkage approach, Wiener filter and Non Local- means algorithm as well. Experimental results showed that the proposed method gives better signal to noise ratio as compared to the previously proposed denoising solutions.

Keywords: Denoising, Edge detection, Lipschitz exponent, Mean square error, Signal Smoothness.

1 Introduction

One of the main problem faced in the field of image processing is that of image denoising, where the goal is to obtain an estimate of the original image from an image that has been contaminated by noise. The two main limitations in any image accuracy are categorized as blur and noise and the main objective of any filtering method is to effectively suppress the noise elements. Not only that, it is of extreme importance to preserve and enhance the edges at the same time. In image processing, the region of abrupt changes contains the most of the useful information about the nature of the image. The region or the points where these changes occurred are termed as an edge. It is possible to study the nature of any edge in terms of singularity or regularity. The singularity is considered to be an important character of an edge, as it refers to the level of discontinuity or interruption present inside the image and the main purpose of the detection of such singular point is to identify the existence, location and size of those singularities. Several methods have been proposed in the past to attain these objectives and to recover the original (noise free) image. Most of these techniques uses averaging filter e.g. the Gaussian smoothing model has been used by Gabor

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 11–18, 2011.

[1], some of these techniques uses anisotropic filtering [2,3] and the neighborhood filtering [4,5], some works in frequency domain e.g Wiener filters [4]. In the past few years, wavelet transform has also been used as a significant tool to denoise the signal [6,7,8]. A brief survey of some of these approaches is given by Buades et al [9].

Traditionally, linear models have been used to extract the noise elements e.g. Gaussian filter as they are computationally less expensive. However, in most of the cases the linear models are not able to preserve sharp edges which are later recognized as a discontinuities in the image. On the other hand, nonlinear models can effectively handle this task (preserve edges) but more often, non linear model are computationally expensive. In the present work, we attempt to propose a non linear model with the very less computational cost to restore image from noisy data samples. The method utilizes data samples and find the best compromise between the data samples and smoothness criteria which ultimately result in giving the denoise signal at a very low computational cost. we have also presented the comparative analysis of the present technique with some of the previously proposed method.

The principle of the proposed technique is given in section 2. Section 3 presents the singular points extraction method. Section 4 explains the overview of restoration method. Comparative analysis is given in section 5 and finally section 6 conclude the work.

2 Principle of the Method

We assume that the given data specify the model:

$$y_{ij} = f(x_{ij}) + \epsilon_{ij} \qquad where \quad i,j = 1,...,n \tag{1}$$

f is the noise free signal, uniformly sampled (e.g., an image) and ϵ_{ij} is the white gaussian noise $N(0,\sigma^2)$. In the present work, the given data will always be an $n \times n$ matrix with $n = 2^N$. The aim of the current work is to estimate the function $F\ :\ (f(x_{ij}))_{i,j=1}^{n}$ with respect to an estimator $\hat{F}$ such that:

$$SNR(dB) = -10log_{10}\frac{\sum_{i,j=1}^{n}(\hat{F}_{i,j} - F_{i,j})^2}{\sum_{i,j=1}^{n} F_{i,j}^2} \tag{2}$$

In order to estimate the function F, the method utilizes the data samples and performs non linear functioning to estimate the best fit. The filtering has been performed on each row and column matrix individually and at the final stage by utilizing filtering in x and y directions yield in fully denoised image.

$$G = \frac{1}{\sqrt{2}}\sqrt{(G_x)^2 + (G_y)^2} \tag{3}$$

where $\frac{1}{\sqrt{2}}$ is the normalizing factor, G_x is the filtered image in horizontal direction and G_y is the filtered image in vertical direction.

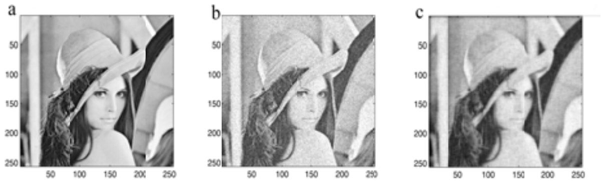


Fig. 1. a) Original Lena image, b) Lena image with "white Gaussian noise" SNR of 15dB, c) Denoised Lena image with proposed Mse-Smooth method

3 Estimation of Singular Points

In this section, the Modulus maxima approach has been applied on signal to identify different signatures in the signal. Mallat proposed a method to compute the singularity of the signal by finding the local maxima. This singularity can be defined in terms of an interval or as a point wise. If the signal has constant behavior in some given interval [a, b] then the singularity is defined in terms of interval. In our work, we focused in computing the singularities at individual points [10].

3.1 Continuous Wavelet Transform

The Morlet-Grossmann definition of the continuous wavelet transform for a 1-dimensional signal $f(x)$, the space of all square integrable functions, is [10]:

$$WT(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(x) \mathrm{y}^* [\frac{(x-b)}{a}] dx \tag{4}$$

Where WT(a,b) is the wavelet coefficient of the function $f(x)$, x is the analyzing wavelet, a (>0) is the scale parameter and b is the position parameter.

The constant $\frac{1}{\sqrt{a}}$ is used to normalise or conserve the energy before and after the transform. It has been explained in that if the wavelet used is the first derivative of a smoothing function, then the wavelet transform $W_1 f(s,x)$ or $W_2 f(s,x)$ is proportional to the first (or the second) derivative of$f(x)$ smoothed by the function wavelet function.

3.2 Lipschitz Exponent

The singularity of the points can be described by computing their Lipschitz exponent. The Lipschitz exponent is a generalized measure of the differentiability of a function

Definition: Supposes n is an integer, $n < \alpha < n+1$, the signal $f(x)$ has Lipschitz a at x_o ,if and only if there exists a constant A and $x_0 > 0$ which, that for the polynomial $P_n(x)$ of n-order, satisfy [10,11,12]:

$$f(x) - p_n(x) \leq A|x - x_0|^{\alpha} \tag{5}$$

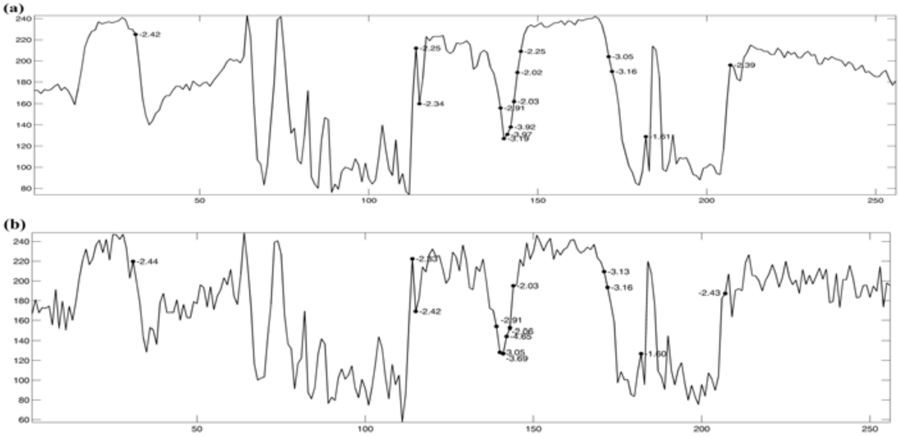


Fig. 2. a) Row No. 125 of Original Lena image, b) Row No. 125 of Noisy Lena image. Modulus maxima are marked as a circle in both images and the corresponding Lipschitz value is written in numerical form.

Where the least upper bound of α at the point of x_0 is defined as the regularity of $f(x)$ at the point of x_0. These Lipschitz exponent can be estimated form the slope of logarithm of modulus maxima lines across logarithm of scale by using above expression. Fig 2. shows the Lipschitz analysis of a single row of original and Noisy Lena image.

4 Reconstruction

The present method proposes restoring by taking into account extracted regular data samples based on their respective Lipschitz exponents as shown in Fig. 3. The method utilizes all sampled points and smoothness (Mse-Smooth) of the signal to estimate the best fit by working in an iterative mode. The method is design for one dimensional signal, therefore it performs non linear filtering on image initially row by row and then column by column. At the final stage, by using expression in given $eq.$3 merging of filtering operation in horizontal and vertical direction results in giving the fully denoised image.

Lets define any one dimensional signal as:

$$y^{k+1} = y^k + \lambda^k(-\frac{\partial C^{MSE,k}}{\partial y}) + \gamma^k(-\frac{\partial C^{MSO,k}}{\partial y}) \tag{6}$$

y denote the one dimensional signal with $l \epsilon 1, ...2^N$ samples. k define the iteration step. C^{MSE} is the mean square error estimation of the restored signal with the original signal (y_o) such that :

$$C^{MSE} = \sum_{l=1}^{N}(y_l^k - y_o^k)^2 \tag{7}$$

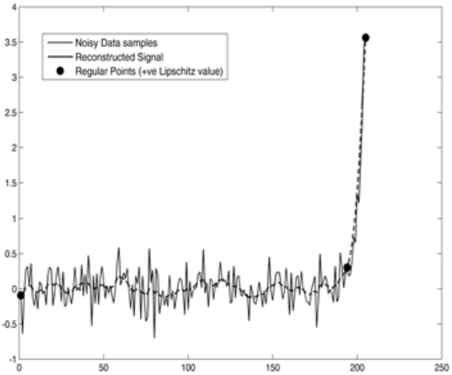


Fig. 3. Signal showing regular points (Lipschitz value $\alpha > 1$) marked as circle. The reconstruction method preserve these regular points. Solid Line = Noisy data samples, Dotted lines = Reconstructed signal.

C^{MSO}define the smoothness of the reconstructed subset signal.

$$C^{MSO} = \sum_{l=1}^{N} y_l^{k''^2} \tag{8}$$

In order to find the value of λ^k and γ^k in *eq*.6, consider the Taylor series expansion and for the given series, in order to find the minimum mean square error we want $f^{'}(x+dx) = 0$ therefore by simplifying, $dx = -\frac{f^{'}(x)}{\frac{d^2 f}{dx^2}}$. we know that: $x^{k+1} = x^k + \lambda dx$. As $x \rightarrow Y^{k+1} \quad and \quad f \rightarrow C^{MSE,k}$ therefore by replacing the variables in *eq* we have :

$$y^{k+1} = y^k + \frac{-C^{MSE,k^{'}}(y)}{\frac{d^2 C^{MSE,k}}{dy^{k^2}}} \tag{9}$$

hence, $\lambda^k = \frac{d^2 C^{MSE,k}}{dy^{k^2}}$ and by simplifying:

$$y^{k+1} = y^k + \lambda^k \left(-\frac{\partial C^{MSE}}{\partial y}\right) \tag{10}$$

Similarly by using the Taylor series in term of smoothing criteria, γ^kcomes out to be:

$$\gamma^k = \frac{d^2 C^{MSO,k}}{dy^{k^2}} \tag{11}$$

By using *eq*.6, each row wise and column wise vector were restored individually and result in giving two $n \times n$ matrices, G_x and G_y respectively. At the last stage, by using mathematical expression given in *eq*.3, the final denoised image has been restored.

5 Results and Discussion

The test image use in this work is Lena 256 × 256 picture. We generated noisy data from clean image by adding pseudorandom numbers (white gaussian noise) resulting in signal to noise ration (SNR) of approximately 15dB (SNR is defined in *eq*.2). Fig.1 shows the corrupted Lena image with white Gaussian noise (15db) and the denoising result with the proposed technique. It can be seen from the figure that, the new method denoised the image reasonably well while keeping the edges preserved. In order to illustrate the amount of the noise in the data and the effects of the denoising, we show in Fig.4, a single row (100) of the image (15dB white Gaussian noise), plotted as a curve. It can be seen from the figure that the method not only smooth the noisy part of the signal but also preserve the sharp edges or transitions to good extent. At the same time, the proposed

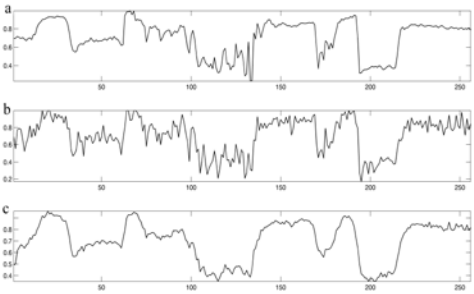


Fig. 4. a) Row 100 of original Lena image, b) Row 100 of noisy Lena image with "white Gaussian noise" SNR of 15dB, c) Row 100 of denoised Lena image with proposed Mse-Smooth method

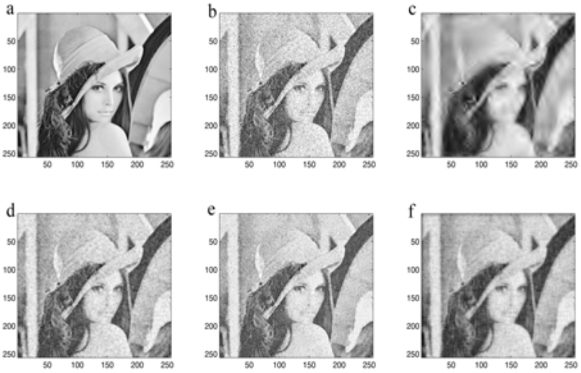


Fig. 5. a) Original Lena image, b) Lena image with "white Gaussian noise" SNR of 15dB, c) Denoised Lena image with Visu shrink method, d) Denoising with Wiener filtering, e) Denoising with Non Local means, f) Denoising with proposed Mse-Smooth method

method performs the whole operations at a very low computational cost which makes it useful for many real time applications (maximum time taken for any experiment was approx 16 sec).

5.1 Comparative Analysis

The human eye is able to decide if the quality of the image has been improved by the denoising method. Fig.5 displays the comparative analysis of the proposed method with other smoothing filters including Non-Local means [13], Wiener filter [4] and classical Visu shrink method [7]. The experiment has been simulated by adding white gaussian noise. It can be seen from the figure that the Visu shrink result in giving good smoothness but at the cost of the edges. Not a single edge has been preserved, this however is not, in the case of NL-means or Wiener filtering. NL-means and Wiener filtering preserved the edges reasonably well, but in this case the noise elements are visible (clearly visible on the background of Lena) and can be seen with the naked eye as well. The proposed (Mse-Smooth) method gave the best compromise of the two (blur and noise artifacts) as shown in Fig.5. Statistical results in terms of SNR (dB) on Lena Image with different types of noise on different method is given Table 1.

Table 1. Statistical results in terms of SNR (dB) of "Lena Image" with different types of noise (M/S is proposed Mean-Smooth method)

Noise Type	Input	NLM	Weiner	Shrink	M/S
Gaussian	16.02	18.12	20.32	17.14	22.56
Speckle	17.20	18.1	23.46	21.03	21.82
Salt/Pepper	14.34	16.13	17.02	17.32	19.23

6 Conclusion

In this work, we have presented a new denoising algorithm, based on the actual noisy image samples. Unlike other existing techniques, we have not considered modeling of an image or noise characteristics in developing the approach. Instead we have estimated the best fit of the signal by utilizing actual noisy data samples and smoothness criteria. At the same time, the proposed non linear image filtering technique works equally well in the presence of signal dependent nature of multiplicative noise in spatial domain. The proposed non linear expression have effectively directed the algorithm in the smoothing operation at a very low computational cost. For images, it is important that edges data should be preserved. The denoising algorithm presented in this work, to a large extent has satisfied the constraint that phase should not be corrupted. The effectiveness of this technique encourages the possibility of improving this approach to preserve the edges to more extent.

References

1. Bruckstein, A., Lindenbaum, M., Fischer, M.: On gabor contribution to image enhancement. pattern recognition. Computer Methods and Programs in Biomedicine 27(1), 1–8 (1994)
2. Malik, J., Perona, P.: Scale space and edge detection using anisotropic diffusion. IEEE Transactions on Pattern Analysis, 629–639 (March 1990)
3. Morel, J.M., Alvarez, L., Lions, P.L.: Image selective smoothing and edge detection by nonlinear diffusion. Journal of Numerical Analysis 29, 845–866 (1992)
4. Yaroslavsky, L.: Digital picture processing - an introduction. Springer, Heidelberg (1985)
5. Manduchi, R., Tomasi, C.: Bilateral filtering for gray and color images. In: Proceedings of the Sixth International Conference on Computer Vision, pp. 839–846 (1998)
6. Donoho, D., Coifman, R.: Translation-invariant de-noising. In: Wavelets and Statistics, pp. 125–150. Springer, Heidelberg (1995)
7. Donoho, D.: De-noising by soft-thresholding. IEEE Transactions on Information Theory 41, 613–627 (1995)
8. Alecu, A., Munteanu, A., Tessens, L., Pizurica, A.: Context adaptive image denoising through modeling of curvelet domain statistics. Journal of Electronic Imaging 17, 033021–17 (2008)
9. Morel, J., Buades, A., Coll, B.: On image denoising methods, Technical Report CMLA (2004)
10. Mallat, S.: A wavelet Tour of Signal Processing. Academic Press, London (1998)
11. Laligant., O., Jalil, B., Fauvet, E.: Noise and artifacts removal utilizing significant features. In: Signal Processing Symposium, Poland, June 8-10 (2011)
12. Laligant., O., Jalil, B., Fauvet, E.: Piecewise image denoising utilizing discontinuous edges. In: The Eleventh International Conference on Pattern Recognition and Information Processing, Minsk, Belarus, May 18-20
13. Morel, J.M., Buades, A., Coll, B.: A non-local algorithm for image denoising. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 60–65 (2005)

Automatic Human Action Recognition in Videos by Graph Embedding

Ehsan Zare Borzeshi, Richard Xu, and Massimo Piccardi

School of Computing and Communications,
Faculty of Engineering and IT University of Technology, Sydney (UTS)
Sydney, Australia
{ezarebor,ydxu,massimo}@it.uts.edu.au
www.inext.uts.edu.au

Abstract. The problem of human action recognition has received increasing attention in recent years for its importance in many applications. Yet, the main limitation of current approaches is that they do not capture well the spatial relationships in the subject performing the action. This paper presents an initial study which uses graphs to represent the actor's shape and *graph embedding* to then convert the graph into a suitable feature vector. In this way, we can benefit from the wide range of statistical classifiers while retaining the strong representational power of graphs. The paper shows that, although the proposed method does not yet achieve accuracy comparable to that of the best existing approaches, the embedded graphs are capable of describing the deformable human shape and its evolution along the time. This confirms the interesting rationale of the approach and its potential for future performance.

Keywords: Graph edit distance, Graph embedding, Object classification.

1 Introduction and Related Work

Human action recognition has been the focus of much recent research for its increasing importance in applications such as video surveillance, human-computer interaction, multimedia and others. Recognising human actions is challenging since actions are complex patterns which take place along the time. Due to the nature of human physiology and the varying environmental constraints, each person performs the same action differently for every instance. More so, different people may perform the same action in a pronouncedly different way in both spatial extent and temporal progression. When translated into feature vectors, actions give place to a probing feature space with very high intra-class variance and low inter-class distance. To mollify this issue, this paper presents an initial study into the possibility of using *graph embedding* for obtaining a more suitable feature set.

Many approaches have been proposed for human action recognition to date, including bag of features [4] [8], dynamic time warping [1], hidden Markov models [21] and conditional random fields [14]. A recent survey has offered a systematic review of these approaches [13]. However, the problem of a suitable feature set which can well encapsulate the deformable shape of the actor is still partially unresolved. Graphs offer

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 19–28, 2011.

a powerful tool to represent structured objects and as such are promising for human action recognition. Ta *et al.* in [19] have recently used graphs for activity recognition. However, to assess the similarity of two instances, they directly compare their graphs which is prone to significant noise. An alternative to the direct comparison of action graphs is offered by graph embedding: in each frame, the graph representing the actor's shape can be converted to a finite set of distances from prototype graphs, and the distance vector then used with conventional statistical classifiers. Graph embedding has been successfully used in the past for fingerprint and optical character recognition [17]. To the best of our knowledge, this is the first work proposing to employ graph embedding for human action recognition. Such an extension is not trivial since feature vectors need to prove action-discriminative along the additional dimension of time. In this paper, we propose to extract spatial feature points from each frame and use them as nodes of a graph describing the actor's shape. With an adequate prototype set, we convert the graph to a set of distances based on the probabilistic graph edit distance (GED) of Neuhaus and Bunke [11]. Probabilistic GED is a sophisticated edit distance capable of learning edit costs directly from a training set and weigh each edit operation individually. The feature vectors of each frame are then composed into a sequence and analysed by means of a conventional sequential classifier. The recognition accuracy that we obtain is not yet comparable with that of the best methods from the literature; however, results show unequivocally that the embedded vectors are capable of representing the human posture as it evolves along the time and set the basis for potential future improvements.

The rest of this paper is organised as follows: Section 2 provides a brief recall of graph embedding. Section 3 describes the methodology proposed by this paper to incorporate graph embedding into an action recognition approach. Section 4 presents and discusses an experimental evaluation of the proposed approach on the challenging KTH action dataset. Finally, we give concluding remarks and a discussion of future work in section 5.

2 A Brief Recall of Graph Embedding

Based on various research studies, different definitions for graphs can be found in the literature. In this work we use an *attributed graph* g represented by $g = (V, E, \alpha, \beta)$ where

- $V = \{1, 2, ..., M\}$ is the vertices (nodes) set, where $M \in \mathbb{N} \cup \{0\}$,
- $E \subseteq (V \times V)$ is the set of edges,
- $\alpha : V \rightarrow L_V$ is a vertex labeling function, and
- $\beta : E \rightarrow L_E$ is a edge labeling function.

Vertex and edge labels are restricted to fixed-size tuples, ($L_V = \mathbb{R}^p$, $L_E = \mathbb{R}^q$, p, $q \in \mathbb{N} \cup \{0\}$).

With a graph-based object representation, the problem of pattern recognition changes to that of graph matching. One of the most widely used methods for error-tolerant graph matching is the graph edit distance (GED). It measures the (dis)similarity of arbitrarily structured and arbitrarily labeled graphs and it is flexible thanks to its ability to cope with any kind of structural errors [5], [3]. The main idea of the graph edit distance is: find the dissimilarity of two graphs by the minimum amount of distortion required to transform one graph into the other [5]. In the first step, the underlying distortion models (or edit operations) are defined as insertion, deletion and substitution for both nodes and edges, $(e_1, e_2, e_3, e_4, e_5, e_6)$. Based on this definition, every graph can be transformed into another by applying a sequence of edit operations (i.e. an edit path). Then, given a set of edit operations and an edit cost function, the dissimilarity of a pair of graphs is defined as the minimum cost edit path that transforms one graph into the other. Let $g_1 = (V_1, E_1, \alpha_1, \beta_1)$ and $g_2 = (V_2, E_2, \alpha_2, \beta_2)$ be two graphs. The graph edit distance of such graphs is defined as:

$$d(g_1, g_2) = \min\,(e_1, ..., e_k) \in E(g_1, g_2) \sum_{i=1}^{k} C(e_i) \tag{1}$$

where $E(g_1, g_2)$ denotes the set of edit paths between two graphs, C denotes the edit cost function and e_i denotes the individual edit operation. Based on (1), the problem of evaluating the structural similarity of graphs is changed into the problem of finding a minimum-cost edit path between two graphs.

Among different methods, the *probabilistic graph edit distance* (P-GED) proposed by Neuhaus and Bunke [10], [11] was chosen to automatically find the cost function from a labeled sample set of graphs. To this aim, the authors represented the structural similarity of two graphs by a learned probability $p(g_1, g_2)$ and defined the dissimilarity measure as:

$$d(g_1, g_2) = -\log p(g_1, g_2) \tag{2}$$

The main advantage of this model is that it learns the costs of edit operations automatically and is able to cope with large sets of graphs with huge distortion between samples of the same class [10], [11].

2.1 Graph Embedding

Graph embedding converts a graph into an n-dimensional real vector. Its motivation is that of trying to take advantage of the rich space of statistical pattern recognition techniques yet retaining the spatial representational power of graphs. Let $G = \{g_1, g_2, ..., g_m\}$ be a set of graphs , $P = \{p_1, p_2, ..., p_n\}$ be a set of prototypes with $m > n$ (detail in subsection 2.2), and d be a (dis)similarity measure (detail in section 2). For graph embedding, dissimilarity d_{ji} of graph $g_j \in G$ to prototype $p_i \in P$ is computed. Then, an n-dimensional vector $(d_{j1}, ..., d_{jn})$ can be achieved through the computation of the n dissimilarities, $d_{j1} = d(g_j, p_1), ..., d_{jn} = d(g_j, p_n)$. As a result of this, any graph $g_j \in G)$ can be transformed into a vector of real numbers.

Formally, the mapping $t^P : G \rightarrow \mathbb{R}^n$ is defined as the following function:

$$t^P(g) \rightarrow (d(g, p_1), ..., d(g, p_n)) \tag{3}$$

where $d(g, p_i)$ is a dissimilarity measure between graph g and prototype p_i [17], [12].

2.2 Prototype Selector

Based on the definition in section 2.1, selecting informative *prototypes* from the underlying graph domain plays a vital role in graph embedding. In other words, in order to have meaningful vectors in the embedding space, a set of selected prototypes $P = \{p_1, p_2, ..., p_n\}$ should be uniformly distributed over the whole graph domain while avoiding redundancies in terms of selection of similar graphs [12], [17], [7].

3 Methodology

Our approach to action recognition is based on i) using graph embedding to create a feature vector of the actor in each frame, ii) concatenating all the feature vectors from the first to the last frame of the action video into a vector sequence, and iii) using a sequential classifier for action classification. As sequential classifier, we have used the well-known hidden Markov model [15]. Moreover, prior to extracting graphs of the actor's shape, we have used a tracker to extract a bounding box of the actor in each frame [2]. Due to limitation in space, we do not describe the tracker and classifier further and focus the next paragraphs on graph construction and embedding.

3.1 Graph Building

A number of SIFT keypoints are extracted within the actor's bounding box in each video frame using the software of Vedaldi and Fulkerson [20]. Based on the chosen threshold, this number typically varies between 5 and 8. Moreover, a Gaussian outlier elimination method is applied on the selected SIFT keypoints to eliminate points which are estimated to be far away from the actor. Example results after these steps are illustrated in figure 1. Then, the location of the remaining SIFT keypoints (x, y) is expressed relatively to the actor's centroid and employed as a node label for an attributed graph describing the actor's shape. In a preliminary study not reported in this paper, we found that graphs with only labelled nodes performed as well as graphs with both labelled nodes and edges and were faster to process. We therefore decide to employ graphs consisting only of labelled nodes.

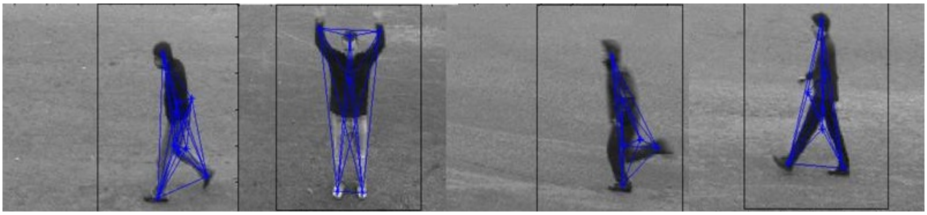


Fig. 1. Bounding box generated by the proposed tracker in the KTH action dataset and final selected SIFT keypoints which are used to build a graph.

3.2 Posture Selection

In order to have a semantic prototype set which could lead to meaningful feature vectors in the embedded space, a number of different reference postures was chosen to describe all of the human shapes in the action dataset. For the dataset at hand, (KTH [18]; details provided in section 4.1), we chose a set of 16 different reference postures across all of the human actions (running, walking, boxing, jogging, hand-waving, hand-clapping). For training purposes, we manually selected a number of different frames varying in scenario (e.g. outdoor, outdoor with different clothes, indoor), action (e.g. hand waving, hand clapping, jogging) and actor (e.g. person01, person25, person12) (see figure 2).

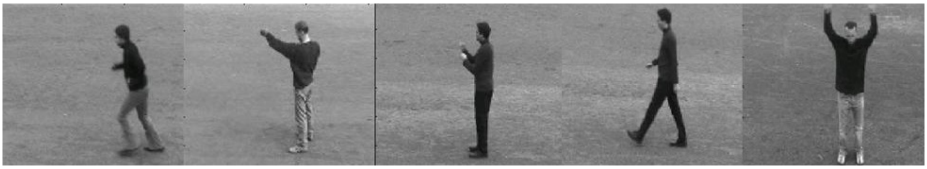


Fig. 2. Examples of selected postures which are used to describe all of the human actions in the KTH action dataset

3.3 Prototype Selection

Among various prototype selection algorithms [17], [12], [7], [16], the *class-wise center prototype selection* (c-cps) method [17] was chosen in this study. With this method, a prototype set $P = \{p_1, ..., p_n, ..., p_N\}$ is found from a class set $C = \{c_1, ..., c_n, ..., c_N\}, N = 16$, with each p_n prototype located in, or near, the centre of class c_n. For selecting the center graph from the sample set of class $c_n = \{g_{n1}, ..., g_{nj}, ..., g_{nN_n}\}$, we choose g_{nj} such that the sum of distances between g_{nj} and all other graphs in c_n is minimal (eq.4) [17].

$$p_n = g_{nj} = \arg \min_{g_{nj} \in c_n} \sum_{g_{ni} \in c_n} d(g_{nj}, g_{ni}) \tag{4}$$

3.4 Feature Vector

The embedding of a graph leads to a 16-dimensional feature vector describing the shape of a single actor in a frame. In order to exploit other available information, we add the displacement between the bounding boxes of two successive frames (which is proportional to the horizontal speed component) and the location of the actor's centroid relative to the bounding box. This leads to an overall 19-dimensional feature vector to describe the shape, motion and location of the actor in each frame (see figure 3).

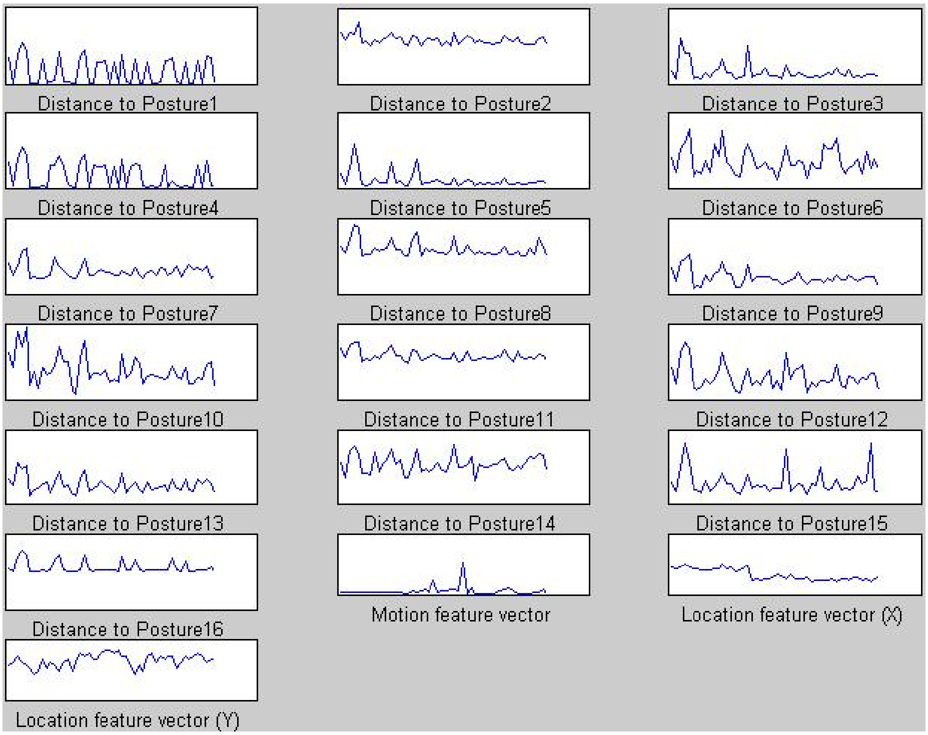


Fig. 3. The time-sequential values of a 19-dimensional feature vector obtained from graph embedding for one action (boxing) performed by one subject in the KTH action dataset

4 Experiments

For the experimental evaluation of our approach, we have chosen a popular dataset, KTH [18], which allows comparison of our results with other, state-of-the-art action recognition methods.

4.1 KTH Action Dataset

The KTH human action dataset contains six different human actions: walking, jogging, running, boxing, hand-waving and hand-clapping, all performed various times over homogeneous backgrounds by 25 different actors in four different scenarios: outdoors, outdoors with zooming, outdoors with different clothing and indoors. This dataset contains 2391 sequences, with each sequence down-sampled to the spatial resolution of 160 × 120 pixels and a length of four seconds on average. While this dataset consists of simplified actions, it is challenging in terms of illumination, camera movements and variable contrasts between the subjects and the background. KTH has been a de-facto benchmark in the last few years and many results are available for comparison.

4.2 Experimental Set-Up and Results

In this section, we evaluate the recognition accuracy of the proposed method. We first evaluate various choices of feature vectors and then compare our approach based on the best feature vector with the state of the art. All of these experiments were performed on a computer with an Intel(R) Core(TM)2 Duo CPU (E8500, 3.16GHz) and 4GB RAM using Matlab R2009b.

Evaluation of the Feature Vectors. The 19-dimensional feature vector described in section 3.4 contains shape, motion and location features jointly. In order to assess the individual contribution of these different types of features, we have conducted experiments with feature vectors containing only shape, motion or location features in isolation. To this aim, we have used *leave one (actor) out cross validation* reporting a correct classification rate (CCR) for each feature vector. It is possible to see that none of the individual type of features was capable of achieving high accuracy in isolation; in all cases, recognition accuracy was below 50% (table 1). However, these features show interesting complementarity: for instance, the motion features report good accuracy in recognising the Jogging class, but a rather low performance on the Boxing class (which is mainly a stationary class). Conversely, the graph-embedded shape features report good accuracy on the Boxing class, but cannot discriminate well between classes such as Jogging and Running where the articulated shape is similar, yet speed of execution varies remarkably. This complementarity is at the basis of the higher performance achieved by the joint vector which jumps to 70.00%, as shown by table 2.

Table 1. The average CCRs on the KTH action dataset based on separate feature vectors for motion, location and shape

validation technique	motion	location	shape
LOOCV-CCR	49.34%	45.67%	47.63%

Comparison to the State of the Art. Accuracy measurements on the KTH database have been performed with different methods by different papers in the literature. For easier comparison, in this section we have used the test approach presented by Schuldt *et al.* in [18]. With this test approach, all sequences are divided into 3 different sets with respect to actors: training (8 actors), validation (8 actors) and test (9 actors). The classifier is then tuned using the first two sets (training and validation sets), and the accuracy on the test set is measured by using the parameters selected on the validation set, without any further tuning. The confusion matrix obtained with the proposed approach is presented in table 3. The overall accuracy is 70.17%, slightly higher than the 70.00% obtained with the leave one out cross validation. This result is not yet comparable with the best accuracies reported in the literature: it is not far from the accuracy reported by *Schuldt et al.* [18], but much lower than that reported by Guo *et al.* in [6] (table 4).

Table 2. Action confusion matrix (%) for the proposed method based on the LOOCV test approach on the KTH action dataset. The average CCR is 70.00%.

	Boxing	Clapping	Waving	Jogging	Running	Walking
Boxing	80	9	8	1	1	1
Clapping	11	59	25	1	2	2
Waving	8	22	66	1	0	3
Jogging	0	0	0	56	21	23
Running	0	0	0	17	74	9
Walking	0	0	0	8	7	85

Table 3. Action confusion matrix (%) for the proposed method based on the Schuldt test approach on the KTH action dataset. The average CCR is 70.17%.

	Boxing	Clapping	Waving	Jogging	Running	Walking
Boxing	92	1	5	0	1	1
Clapping	18	62	17	1	1	1
Waving	9	35	55	0	0	1
Jogging	1	0	0	56	18	25
Running	1	0	0	14	66	19
Walking	0	0	0	5	5	90

Table 4. Average class accuracy on the KTH action dataset

Method	**Ours**	Schuldt *et al.*[18]	Dollar *et al.*[4]	Laptev *et al.*[9]	Guo *et al.*[6]
LOOCV-CCR	**70.00%**	-	80%	-	98.47%
Schuldt-CCR	**70.17%**	71.70%	-	91.80%	97.40%

4.3 Discussion

In section 4.2 we have showed our initial experimental results from the application of graph building and embedding to human action recognition. Despite the good posture discrimination provided by P-GED (not reported quantitatively here for reasons of space), the overall action recognition accuracy on the KTH dataset is not yet very high. Based on our judgment, the main difficulty faced by the proposed approach is the extraction of a reliable set of keypoints in each frame. Due to noise and variable appearance, the extracted set changes significantly over the frames. Another possible limitation is the limited accuracy of the employed classifier (HMM). However, we believe that the work conducted to date already provides evidence that the features obtained by graph embedding are capable of encoding the actor's shape to a significant extent.

5 Conclusions and Future Work

In this paper, we have presented a novel approach for human action recognition based on graph embedding. To this aim, an attributed graph is used to represent the actor's shape in each frame and then graph embedding is used to convert the graph into a feature vector so as to have access to the wide range of current classification methods. Although our method does not yet match the accuracy of existing approaches, it generates a novel methodology for human action recognition based on graph embedding and may outperform existing methods in the future. With reference to limitations discussed in section 4.3, we plan to further investigate other keypoint sets to improve the stability of the graph-based representation along the frame sequence and employ different classification methods for the classification stage.

Acknowledgments. The authors wish to thank the Australian Research Council and its industry partners that have partially supported this work under the Linkage Project funding scheme - grant LP 0990135 “Airport of Future“.

References

1. Blackburn, J., Ribeiro, E.: Human motion recognition using isomap and dynamic time warping. In: Elgammal, A., Rosenhahn, B., Klette, R. (eds.) Human Motion 2007. LNCS, vol. 4814, pp. 285–298. Springer, Heidelberg (2007)
2. Chen, T., Haussecker, H., Bovyrin, A., Belenov, R., Rodyushkin, K., Kuranov, A., Eruhimov, V.: Computer vision workload analysis: case study of video surveillance systems. Intel Technology Journal 9(2), 109–118 (2005)
3. Conte, D., Foggia, P., Sansone, C., Vento, M.: Thirty years of graph matching in pattern recognition. International Journal of Pattern Recognition and Artificial Intelligence 18(3), 265–298 (2004)
4. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72. IEEE, Los Alamitos (2006)
5. Gao, X., Xiao, B., Tao, D., Li, X.: A survey of graph edit distance. Pattern Analysis & Applications 13(1), 113–129 (2010)
6. Guo, K., Ishwar, P., Konrad, J.: Action Recognition Using Sparse Representation on Covariance Manifolds of Optical Flow
7. Hjaltason, G., Samet, H.: Properties of embedding methods for similarity searching in metric spaces. IEEE Transactions on Pattern Analysis and Machine Intelligence, 530–549 (2003)
8. Laptev, I.: On space-time interest points. International Journal of Computer Vision 64(2), 107–123 (2005)
9. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE, Los Alamitos (2008)
10. Neuhaus, M., Bunke, H.: A probabilistic approach to learning costs for graph edit distance. In: Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, vol. 3, pp. 389–393. IEEE, Los Alamitos (2004)
11. Neuhaus, M., Bunke, H.: Automatic learning of cost functions for graph edit distance. Information Sciences 177(1), 239–247 (2007)

12. Pekalska, E., Duin, R.: The dissimilarity representation for pattern recognition: foundations and applications. World Scientific Pub. Co. Inc., Singapore (2005)
13. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28(6), 976–990 (2010)
14. Quattoni, A., Wang, S., Morency, L.P., Collins, M., Darrell, T., Csail., M.: Hidden-state conditional random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2007)
15. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
16. Riesen, K., Bunke, H.: Graph classification by means of Lipschitz embedding. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 39(6), 1472–1483 (2009)
17. Riesen, K., Neuhaus, M., Bunke, H.: Graph embedding in vector spaces by means of prototype selection. In: Proceedings of the 6th IAPR-TC-15 International Conference on Graph-Based Representations in Pattern Recognition, pp. 383–393. Springer, Heidelberg (2007)
18. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, vol. 3 (2004)
19. Ta, A.-P., Wolf, C., Lavoue, G., Baskurt, A.: Recognizing and localizing individual activities through graph matching, pp. 196–203. IEEE Computer Society, Los Alamitos (2010)
20. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms (2008)
21. Yamato, J., Ohya, J., Ishii, K.: Recognizing human action in time-sequential images using hidden Markov model. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1992. Proceedings CVPR 1992, pp. 379–385 (1992)

Human Action Recognition by Extracting Features from Negative Space

Shah Atiqur Rahman[1], M.K.H. Leung[2], and Siu-Yeung Cho[1]

[1] School of Computer Engineering, Nanyang Technological University, Singapore 639798
shah0018@ntu.edu.sg, davidcho@pmail.ntu.edu.sg
[2] FICT, Universiti Tunku Abdul Rahman (Kampar), Malaysia
asmkleung@gmail.com

Abstract. A region based technique is proposed here to recognize human actions where features are extracted from the surrounding regions of a human silhouette termed as negative space. Negative space has the ability to describe poses as good as the positive spaces (i.e. silhouette based methods) with the advantage of describing poses by simple shapes. Moreover, it can be combined with silhouette based methods to make an improved system in terms of accuracy and computational costs. Main contributions in this paper are two folded: proposed a method to isolate and discard long shadows from segmented binary images, and generalize the idea of negative space to work under viewpoint changes. The system consists of hierarchical processing of background segmentation, shadow elimination, speed calculation, region partitioning, shape based feature extraction and sequence matching by Dynamic Time Warping. The recognition accuracy of our system for Weizmann dataset is 100% and for KTH dataset is 95.49% which are comparable with state-of-the-art methods.

Keywords: Human action recognition, Negative space, Silhouette, Dynamic time warping, complex activity, fuzzy function.

1 Introduction

In the field of computer vision, human action recognition is an attractive research topic due to its application area and challenging nature of the problem. Cluttered background, camera motion, occlusion, shadows, viewing angle changes, and geometric and photometric variances are the main challenges of human action recognition. Application of human action recognition includes virtual reality, games, video indexing, teleconferencing, advance user interface, video surveillance etc.

Despite the fact that good results were achieved by traditional action recognition approaches, they still have some limitations [1, 2]. Tracking based methods [3] suffers from self-occlusions, change of appearance, and problems of re-initialization. Methods based on key frames or eigen-shapes of silhouettes [4] do not have motion information which is an important cue for some actions. Local features are extracted in bags-of-words methods [5], lacking temporal co-relation between frames. Optical flow based techniques [6] face difficulties in case of aperture problems, smooth surfaces, and discontinuities.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 29–39, 2011.

Since region based methods are relatively robust to noise [2], we proposed a region based approach which extract features from the surrounding regions (negative space) of the silhouette rather than the silhouette itself (positive space) [7]. Negative spaces have the ability to describe poses as good as the positive space with the advantage of natural partitioning of negative space into regions of simple shapes. This approach also performed well in case of partial occlusion and small shadows [7]. However, the system could not show good performance under viewing angle change and long shadows. Here, we extend the idea of negative space to recognize actions in case of viewpoint change by modifying the computation of motion feature and incorporating different viewing angle model data. Moreover, we also propose a method to handle long shadows in segmented images which is one of the major challenges of human action recognition.

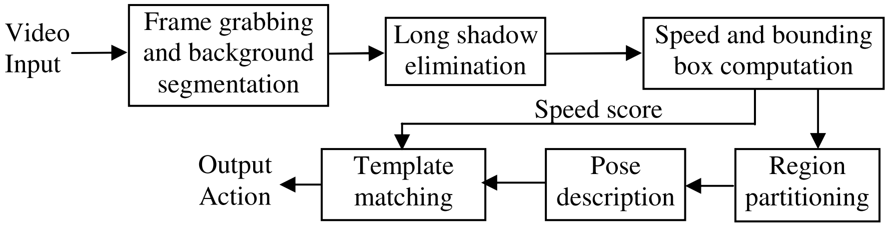


Fig. 1. Block diagram of the system

2 The Proposed System

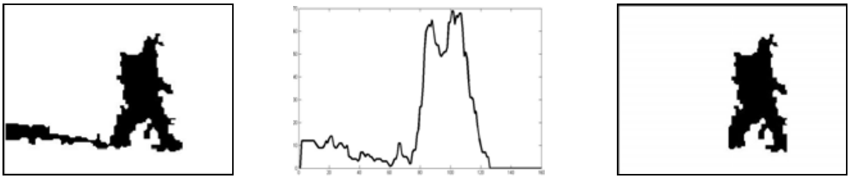
Our system block diagram is shown in Fig. 1. Input of our system is a video containing single person performing an action. Multi-person activity recognition is left as the future work of current study. The input video is first background segmented which is done by Li et al. [8] algorithm in our system. Long shadows are discarded from segmented images in *long shadow elimination* step. Speeds of the human body are computed and negative spaces are captured in the next step. Complex regions are partitioned into simple regions in *region partitioning* step. Positional and shape based features are extracted in *pose description* step to describe each pose. Finally, pose sequences are matched by Dynamic Time Warping (DTW) and input action is recognized by Nearest Neighbor classifier based on the speed and DTW score.

2.1 Long Shadow Elimination

Some system assumes that the shadow is discarded by the segmentation process [4], otherwise those system performances would be degraded. In our case, during segmentation shadow is not discarded which implies our input segmented image may contain long shadows. Negative space processing has the advantage of recognizing actions effectively in presence of small shadow and partial occlusion [7], but in case of long shadow it may fail to recognize. Long shadows have the characteristics that they are connected to the lower part of the body and are projected on the ground (Fig. 2(a)). Motivating from these characteristics, a histogram based analysis is proposed to isolate shadows from the body and then discard it. Number of bins in the histogram is same as the number of columns and frequency of each bin is the pixel count of foreground pixels along the corresponding column (Fig. 2(b)). Empirically we found that the lower foreground pixels (lower than 10% of silhouette height) of columns (bins) that satisfy equation (1) belong to the shadows.

$$\frac{|c_i|}{\max\limits_{k=1}^{n} |c_k|} \leq 0.20 \tag{1}$$

In (1), $|c_i|$ is the pixel count of foreground pixels of the i^{th} column and n is the number of columns in the input image. Hence, we remove lower pixels of those columns and then take the biggest blob from the image. Eventually shadows are discarded from the input image (Fig. 2 (c)).



(a) (b) (c)

Fig. 2. Discarding long shadow. (a) Segmented image with shadow. (b) Histogram of foreground pixels of (a). (c) After discarding the shadow.

2.2 Speed and Bounding Box Computation

For human action recognition, the speed of a person is important cues since different actions are performed in different speeds (e.g. running is performed in a faster speed than walking). Speed of a person can be calculated as

$$hor_{sp} = \frac{ds}{dt} = \frac{|x_i - x_{i-1}|}{1/fr\_rate} = |x_i - x_{i-1}| \times fr\_rate \tag{2}$$

where x_i is the X-axis coordinate of the centroid of human body of i^{th} frame and *fr_rate* is the frame rate of the sequence. To remove the scaling effect we normalize equation (2) by the height of the bounding box of the person. Expressing with respect to the average value we have

$$hor_{sp} = \frac{t\_hor_{disp}/(n-1) \times fr\_rate}{t\_height/(n-1)} = \frac{t\_hor_{disp} \times fr\_rate}{t\_height} \tag{3}$$

where $t\_hor_{disp} = \sum_{i=2}^{n} |x_i - x_{i-1}|$, n is the total number of frames in the sequence and *t_height* is the sum of all bounding box height excluding the first frame. Equation (3) can calculate speed effectively where the action is performed in parallel to the image plane but in case of displacement actions (e.g. walk), not performed in parallel with the image plane (i.e. viewpoint is changed), it calculates the observed speed rather than the actual speed of the body (Fig. 3). In Fig. 3 action is performed in a plane which makes an angle (δ) with the image plane. To calculate the actual speed in this scenario, we need to compute the actual displacement of the person which can be calculated by equation (4) (Fig. 3).

$$t\_hor_{disp} = t\_obs_{disp} / \cos\delta \tag{4}$$

Here δ is calculated as [9]

$$\delta = \frac{1}{2}\tan^{-1}\left[\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}\right] \tag{5}$$

where, μ_{ij} is the i, j^{th} order centralized moment of the centroids (the black dots in Fig. 3) of the human silhouette in the sequence. Hence, the actual speed with viewpoint normalization is

$$ac\_hor_{sp} = \frac{t\_obs_{disp} / \cos\delta \times fr\_rate}{t\_height} = \frac{hor_{sp}}{\cos\delta} \tag{6}$$

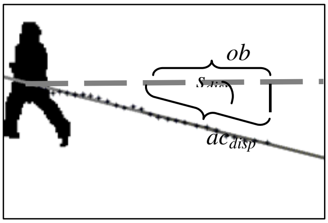


Fig. 3. Viewpoint normalization. Dots are silhouette centroid of each frame, solid line is major principle axis along the centroids, dashed line is the X-axis, δ is the angle between action performing plane and image plane. obs_{disp}, and ac_{disp} are observed and actual displacement respectively.

During speed calculation, we first calculate the speed hor_{sp}, using equation (3). If the human body has significant movement (e.g. walk, run), we apply equation (6) to normalize the speed with respect to viewpoint change and compute the actual speed ac_hor_{sp}. Otherwise (e.g. waving, clapping) ac_hor_{sp} is same as hor_{sp} since for non-moving actions viewpoint change does not affect the speed calculation.

Next an upright bounding box is cut containing the human body to capture negative space regions. Human can perform action in both directions (i.e. moving to the left or right of the image frame). To make the computation easier, we alter all moving action sequences (e.g. walk, run) into one direction (i.e. move to the left of the image frame) by flipping pixels of all the bounding boxes moving left to right about the Y-axis. A movement is seen as from left to right, if the X-coordinate of human body centroid increases over time. For non-moving actions, if the limbs movement is asymmetric (e.g. box), we employed two sets of training data for a single sequence: the flipped and non-flipped images of the sequence, whereas for symmetric movement actions (e.g. handclapping) we employed the training data as it is.

2.3 Region Partitioning

Same pose of same person but captured at different time may not share same number of regions due to the continuous movement of the body as shown in Fig. 4, where pose 4(a) and 4(b) are taken from same pose group but they do not share same number of regions. To overcome the situation and simplify the matching process, region partitioning is applied. We employed the same region partitioning technique as in [7] where for each region, peninsula growing from the silhouette is identified by line growing process. If the peninsula is valid for partition, which is identified by protrusive measure of three distances (Fig. 4(c)), the region is partitioned into two by the tip of the peninsula (Fig. 4(d), 4(e)).

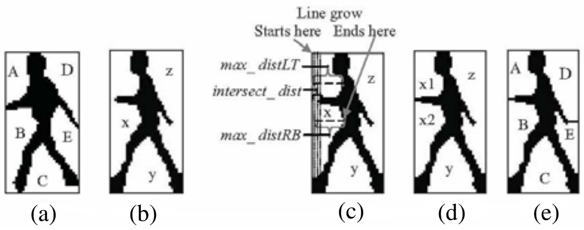


Fig. 4. Region partitioning scenarios. (a) No partition is needed, (b) partition is desired (c) partitioning measures taken for region 'x' of (b), (d) partition output of region 'x'. (e) Final partition output of (b).

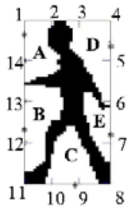


Fig. 5. Positional feature. Numbers represent the anchoring points and letters represent the region, '*' represents the mid-point for each region.

2.4 Pose Description and Matching

Two types of features are extracted to describe the poses: positional feature and shape based features which describe the location and the shape of each negative space region respectively.

2.4.1 Positional Feature

To define the location of a region, we label the bounding box with 14 anchoring points (Fig. 5). For each region, mid-point on the side of the bounding box is computed ('*' in Fig. 5) and the region is assigned a positional label with respect to the nearest anchoring point from that mid-point. For example, anchoring points for regions A, B, C, D and E are points 1, 12, 9, 5 and 7 respectively.

2.4.2 Region Based Features

We extracted simple region based features to describe triangle or quadrangle since negative space regions can be approximated by those shapes [7]. Our shape based features are area, orientation, eccentricity, rectangularity, horizontal and vertical side lengths of bounding box included in a region.

2.4.3 Distance between Poses

Matching of regions for two similar poses raises the need to shift anchor point, e.g. region ‘D’ of Fig. 5 can be assigned to anchor point 6 or 4 instead of 5. In the matching process, we allow the mid point (‘*’) to move at most once to its left or right neighboring anchor point. Hence, we need to develop a distance metric which may allow the region to shift at most one position without any penalty and calculate the distance between poses effectively. This could be done by similar technique used in [7], where a matrix PM is constructed as equation (7)

$$PM\,(i,j)=\begin{cases} r_{dist}\,(v1(i),v2(j)) & \textit{if}\ \ |i-j|\le 1\ \ \textit{or}\ \ |i-j|=13 \\ \infty & \textit{otherwise} \end{cases} \tag{7}$$

where *i, j=1 to14*, *v1* and *v2* are two pose vectors. r_{dist} is defined as

$$r_{dist}(v1(i),v2(j))=\sqrt{\sum_{k=1}^{6}\begin{cases} (v1(i)_k-v2(j)_k)^2 & k\neq 1 \\ (|v1(i)_k-v2(j)_k|/0.5)^2 & k=1\,\&\,|v1(i)_k-v2(j)_k|\le 0.5 \\ ((1-|v1(i)_k-v2(j)_k|)/0.5)^2 & \text{Otherwise} \end{cases}} \tag{8}$$

where k is the index variable of 6 features for each region with orientation being $v1(i)_1$ or $v2(j)_1$. The maximum orientation difference is π/2.

Then algorithm 2.4.1 is applied to calculate the distance between two poses.

Algorithm 2.4.1. *Function f_dist(PM)*

```
Begin
  pose_d=0;
  m_ele=MIN(PM);
  while m_ele≠INF
      [r c]=POSITION(m_ele,PM);
      pose_d=pose_d+m_ele;
      assign INF to all elements of row r in PM
      assign INF to all elements of column c in PM
      if r≠c  //anchor point is shifted
       pose_d=pose_d+PM(c,r);
         assign INF to all elements of row c in PM
         assign INF to all elements of column r in PM
      end
      m_ele=MIN(PM);
  end
  return pose_d
end
```

when *INF=∞*, *MIN(X)* returns the minimum element in matrix *X* and *POSITION(j,X)* returns the location of *j* inside matrix *X* in terms of row and column. If there are multiple *j*, return the location of *j* with lowest row and column values.

2.5 Distance between Sequences

Since the temporal duration of even same type of actions can be different, time warping should be applied to determine the distance between two sequences. We employed DTW algorithm for this purpose. DTW finds a minimized mapping path in terms of pose distance according to a recurrence relation (equation (9)) with respect to some constraints (e.g. relaxed end point, slope constraints) which are same as [7].

$$D(i,j) = d(i,j) + \min\begin{cases} D(i-1,j) + w_v(i,j) \\ D(i-1,j-1) \\ D(i,j-1) + w_h(i,j) \end{cases} \tag{9}$$

$$\textit{Initialization: } D(1,t)=d(1,t),\ D(i,1)=\infty$$

where D is the DTW matrix, $i=2$ *to* n, $j=2$ *to* m, $t=1$ *to* m, $d(i,j)$ being the distance between poses calculated as in previous section, m and n are number of poses in input and model sequences respectively, w_h and w_v are the slope constraints: $w_{h(v)=}cons\_m_{h(v)}$ if $cons\_m_{h(v)}>2$ otherwise 0 ($cons\_m_{h(v)}$ is consecutive move of warping path in horizontal (vertical) direction).

2.5.1 Doubly Matching

We perform doubly matching scheme (Fig. 6) where two matching is done consecutively since, certain action type, e.g. bending, can contain two parts with one part having perfect match with another short action type, e.g. place jumping. The 1st match is determined by DTW matching from model to test sequence which can occur in any position of test sequence. Let k number of test frames are matched. The 2nd match is determined by performing DTW matching on the subsequent *1.1k* frames right after the first match if later part is larger than the former part (Fig. 6) otherwise matching is done on the preceding *1.1k* frames. For the speed variation within a sequence, we allowed *1.1k* frames for the 2nd match. Average of the two matching scores is taken as the distance between the model and test sequence. For recognition, DTW matching score is computed between test sequence and all the model sequences. Next, speed score is added to the individual matching score and then Nearest Neighbor classifier is employed to recognize the action.

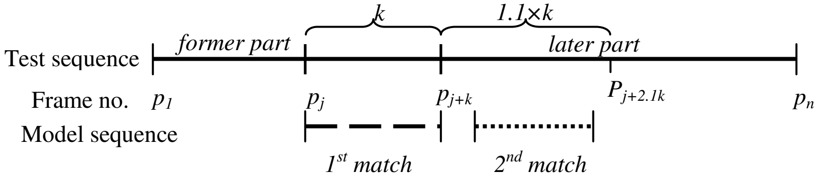


Fig. 6. Matching process of model sequence with test sequence

3 Experimental Results

The proposed system is evaluated by two publicly available datasets: Weizmann human action dataset [10] and KTH action dataset [11].

Weizmann dataset: In this dataset there are 9 persons performing 10 actions. Background segmented images are provided by the author of the dataset and we employed those in our system. Since, the number of video sequence is low, we divide each training sequence into sub-sequences containing only one cycle of an action and we used these sub-sequences as our model sequence. We employed leave-one-out (LOO) testing scheme as most of the other systems used this scheme.

KTH dataset: This dataset is more challenging than Weizmann dataset due to considerable amount of camera movement, long shadow, different clothing and scale and viewpoint variations. There are 25 persons performing 6 actions in 4 different scenarios. Each video is sub-divided into 4 sequences and there are 2391 sequences in the dataset. To extract the silhouette we employ the algorithm by Li et al. [8] without discarding the shadows. Some methods first cut a bounding box either manually [12] or empirically [13] and then perform background segmentation on the bounding box image to reduce noise but in our system we directly feed the input sequence to the segmentation algorithm which means we have less dependency on segmentation process. Since the number of video sequences is high enough, we took only one cycle of an action from each video to avoid unnecessary calculation. Previous systems treat this dataset either as a single dataset (all scenarios in one) or as four different datasets (individual scenarios are treated as a separate dataset, trained and tested separately). Our system is evaluated on both this settings. As a testing scheme some system used Leave-one-person-out (LOO) scheme. Others used split based methods where the

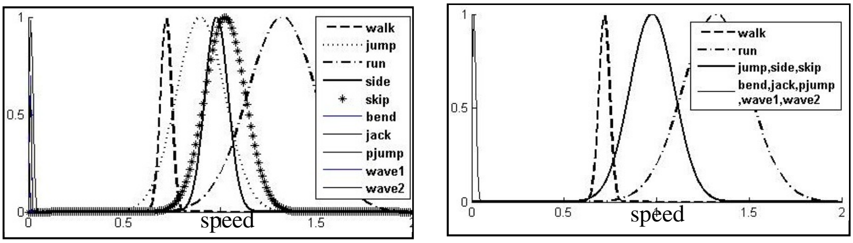


Fig. 7. Membership functions for Weizmann dataset. (a) All 10 membership functions, (b) the final membership functions after merging.

Table 1. Accuracy of different methods for Weizmann dataset

Method	Without skip (%)	With skip (%)
Our Method	**100**	**100**
Fathi et al. [14]	100	100
Ikizler et al. [13]	100	N/A
Kellokumpo et al. [12]	N/A	98.9
Gorelick et al. [10]	N/A	97.8
Lucena et al. [6]	N/A	98.9

Table 2. Accuracy of different methods for KTH dataset

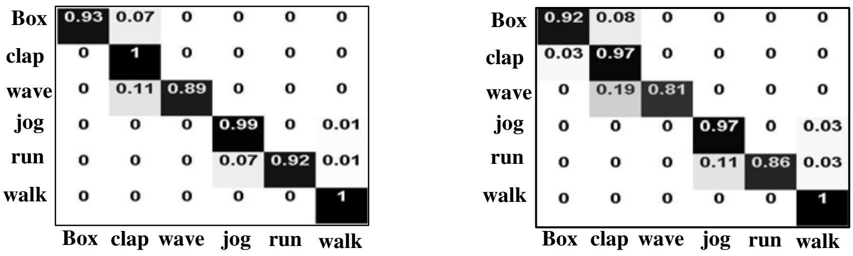
Method	LOO/ Split	Recognition rate (%)	
		Average of all scenes	All scenes in one
Our method	**LOO**	**94.00**	**95.49**
Our method	**Split**	**91.67**	**92.13**
Lin et al. [15]	LOO	95.77	93.43
Kellokumpo et al. [12]	LOO	N/A	95.4
Schindler et al. [16]	Split	90.72	92.7
Fathi et al. [14]	Split	N/A	90.5
Ikizler et al. [13]	Split	N/A	89.4

video sequences are split into training and testing data as [11]. We report results for both testing scheme.

To calculate the speed score, fuzzy membership functions are generated for each type of actions from the training data. Gaussian membership function is chosen as it requires only two parameters (mean and standard deviation) which can be calculated from the training data of each action type. To avoid noisy data, we truncate 5% of data from both highest and lowest boundary, in terms of speed, from each action type. If one membership function is overlapped with another membership function more than 60% of its area, both functions are replaced by a new membership function which is constructed by employing the data of overlapped action types. This merging process goes on until no new membership function is generated. One example is shown in Fig. 7 for Weizmann dataset. Fuzzy functions for KTH dataset is generated by same technique.

Let an input sequence speed is sp_{input} and it is matched with a model of action type 'walk', then the speed score S_C of the input sequence is likelihood score of 'walk' membership function for horizontal speed sp_{input}. Then DTW score is added with speed score by subtracting S_C from 1 since S_C is similarity measure and DTW score is a dissimilarity measure.

Comparison of our system with others for Weizmann dataset is shown in Table 1. Some author evaluated their system with a subset of this dataset (discarding 'skip' action type) which is indicated in Table 1. Our system achieved 100% accuracy on this dataset. Fathi et al. [14] method also achieved perfect accuracy on full dataset but their system requires some parameters to be manually initialized.



	Box	clap	wave	jog	run	walk
Box	0.93	0.07	0	0	0	0
clap	0	1	0	0	0	0
wave	0	0.11	0.89	0	0	0
jog	0	0	0	0.99	0	0.01
run	0	0	0	0.07	0.92	0.01
walk	0	0	0	0	0	1

	Box	clap	wave	jog	run	walk
Box	0.92	0.08	0	0	0	0
clap	0.03	0.97	0	0	0	0
wave	0	0.19	0.81	0	0	0
jog	0	0	0	0.97	0	0.03
run	0	0	0	0.11	0.86	0.03
walk	0	0	0	0	0	1

Fig. 8. Confusion matrix of our system for KTH dataset. (a) leave-one-out scheme (b) split based scheme. For both cases all scenes are taken as a single dataset.

Comparison of our method for different settings with other methods for KTH dataset is shown in Table 2. Our accuracy for this dataset is 95.49% for LOO and 92.13% for split testing scheme which indicates that the amount of training data does not affect too much in our system. Our system performance is better than others for 'all scenes in one'. In case of 'average of all scenes' our system accuracy outperforms most of the other systems. Lin et al. [15] method achieved highest accuracy in this case which is comparable to ours. Lin et al. method is a prototype (key frame) based method which described poses by combination of shape descriptor (describe shape of the silhouette) and motion descriptor (describe motion of different limbs). They need 256-dimensional shape descriptor to describe the shape of pose whereas our pose descriptor is 84-dimensional. Moreover, their system did not employ global motion (speed) of the body to distinguish actions which could significantly improve the accuracy. Hence, by combining Lin et al. [15] method (positive space) with our method (negative space) including speed feature, a new improved system could be obtained in terms of accuracy and computational cost as well. Confusion matrices for both testing scheme is shown in Fig. 8 for 'all-scene-in-one' setting. Most of the misclassification occurs due to the noisy segmentation of background which indicates that if the segmentation process was controlled like other methods [12], our accuracy could be further improved.

4 Conclusion

An action recognition system is proposed here which works on negative space. Our earlier work shows that negative space methods are robust to noise, partial occlusion, small shadows, clothing variations etc [7]. Here, we extend the negative space idea to work under viewing angle change. Moreover, a technique is proposed to discard long shadows from segmented binary images. Additionally, a method is proposed to generate fuzzy membership functions automatically from the training data to calculate speed effectively. Our System accuracy is comparable with the state-of-the-arts methods as shown in experimental results. Further, it is theoretically stated that negative space methods could be combined with positive space methods to obtain an improved system. However, our long shadow detection algorithm may fail in case of multiple shadows in one side of the body (e.g. players' shadows in a football match). Also, our system may not able to recognize actions where the action performing plane and image plane are nearly perpendicular to each other. Nevertheless, these could be overcome when it is combined with positive space based pose description.

References

1. Poppe, R.: A Survey on Vision-based Human Action Recognition. Image and Vision Computing 28(6), 976–990 (2010)
2. Wang, L., Hu, W., Tan, T.: Recent Developments in Human Motion Analysis. Pattern Recognition 36(3), 585–601 (2003)
3. Ikizler, N., Forsyth, D.A.: Searching for Complex Human Activities with No Visual Examples. IJCV 80(3), 337–357 (2008)
4. Diaf, A., Ksantini, R., Boufama, B., Benlamri, R.: A novel human motion recognition method based on eigenspace. In: Campilho, A., Kamel, M. (eds.) ICIAR 2010. LNCS, vol. 6111, pp. 167–175. Springer, Heidelberg (2010)
5. Wang, X., Ma, X., Grimson, W.E.: Unsupervised Activity Perception in Crowded and Complicated Scenes Using Hierarchical Bayesian Models. PAMI 31(3), 539–555 (2009)
6. Lucena, M., de la Blanca, N.P., Fuertes, J.M., Marín-Jiménez, M.: Human action recognition using optical flow accumulated local histograms. In: Araujo, H., Mendonça, A.M., Pinho, A.J., Torres, M.I. (eds.) IbPRIA 2009. LNCS, vol. 5524, pp. 32–39. Springer, Heidelberg (2009)
7. Rahman, S.A., Li, L., Leung, M.K.H.: Human Action Recognition by Negative Space Analysis. In: Cyberworlds (CW), pp. 354–359 (2010)

8. Li, L., Huang, W., Gu, I.Y.-H., Qi, T.: Statistical Modeling of Complex Backgrounds for Foreground Object Detection. Image Processing 13(11), 1459–1472 (2004)
9. Prokop, R.J., Reeves, A.P.: A survey of moment-based techniques for unoccluded object representation and recognition. CVGIP 54(5), 438–460 (1992)
10. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as Space-Time Shapes. PAMI 29(12), 2247–2253 (2007)
11. Schuldt, C., Laptev, I., Caputo, B.: Recognizing Human Actions: A Local SVM Approach. In: ICPR, pp. 32–36 (2004)
12. Kellokumpu, V., Zhao, G., Pietikinen, M.: Dynamic Textures for Human Movement Recognition. In: Int. Conference on Image and Video Retrieval, pp. 470–476 (2010)
13. Ikizler, N., Duygulu, P.: Histogram of Oriented Rectangles: A New Pose Descriptor for Human Action Recognition. Image and Vision Computing 27(10), 1515–1526 (2009)
14. Fathi, A., Mori, G.: Action Recognition by Learning Mid-level Motion Features. In: CVPR, pp. 1–8 (2008)
15. Lin, Z., Jiang, Z., Davis, L.S.: Recognizing Actions by Shape-Motion Prototype Trees. In: ICCV, pp. 444–451 (2009)
16. Schindler, K., van Gool, L.: Action Snippets: How Many Frames Does Human Action Recognition Require? In: CVPR, pp. 1–8 (2008)

Edge-Directed Image Interpolation Using Color Gradient Information

Andrey Krylov and Andrey Nasonov[⋆]

Laboratory of Mathematical Methods of Image Processing,
Faculty of Computational Mathematics and Cybernetics,
Lomonosov Moscow State University,
119991, Leninskie gory, Moscow, Russia
{kryl,nasonov}@cs.msu.ru
http://imaging.cs.msu.ru/

Abstract. Image resampling method using color edge-directed interpolation has been developed. It uses color image gradient to perform the interpolation across image gradient rather than along image gradient. The developed combined method takes color low resolution image and grayscale high resolution image obtained by a non-linear image resampling method as an input. It includes consecutive calculation stages for high resolution color gradient, for high resolution color information interpolation and finally for high resolution color image assembling.

The concept of color basic edges is used to analyze the results of color image resampling. Color basic edge points metric was suggested and used to show the effectiveness of the proposed image interpolation method.

Keywords: color image upsampling, gradient based interpolation.

1 Introduction

Image interpolation is as a key part of many image processing algorithms. For example, image interpolation is performed when a low-resolution video is shown on a high-resolution display. Preserving edge information is the main problem of image interpolation algorithms.

Color image interpolation is generally used in image demosaicing [8] while most of existing approaches in color image upsampling are reduced to grayscale image interpolation by decomposing the image into individual color components and processing these components independently. There exists a large variety of edge-directed image interpolation algorithms [6,7,12], but most of them do not take into account color edges. If the image is converted to a color model with separated luminance (brightness) and chrominance (color) components like YUV, the attention is usually paid to image luminance only while chrominance components are interpolated using simple methods like bilinear or bicubic interpolation.

[⋆] The work was supported by Russian Foundation for Basic Research grant 10-01-00535.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 40–49, 2011.

This approach does not lead to significant degradation of the perceptual image quality because the sensitivity of the human perception to the change of image intensity is higher than to the change of image color. But there is a small portion of edges which are strong in chrominance components and weak in the luminance component. In [4] it was shown that about 10 percent of image edges are not detected in the luminance component. Low quality of interpolation of these edges may be annoying.

In this paper, we present a method to improve the interpolation of chrominance components using multichannel gradient. The concept of color basic edges is used to find the edges where involving color information can significantly improve the quality.

2 Gradient Based Grayscale Image Interpolation

We use gradient based method for the interpolation of single component images. The idea of gradient methods is to use different interpolation kernels depending on image gradient [2]. We calculate the value of pixel being interpolated as a weighted sum of pixels laying along the normal to the gradient in the interpolated pixel.

In our work, we consider linear interpolation method [1]

$$f(x,y) = \frac{\sum\limits_{i,j} u_{i,j} K\left(\sqrt{(x-i)^2+(y-j)^2}\right)}{\sum\limits_{i,j} K\left(\sqrt{(x-i)^2+(y-j)^2}\right)}, \tag{1}$$

where $u_{i,j}$ is the low-resolution image, $f(x,y)$ is the interpolated image, $K(t)$ is the interpolation kernel.

To construct the edge adaptive gradient image interpolation algorithm based on the linear interpolation method, we stretch the interpolation kernel in the normal direction to the image gradient:

$$f(x,y) = \frac{\sum\limits_{i,j} u_{i,j} K\left(\sqrt{(x')^2+(\frac{1}{\sigma_{x,y}}y')^2}\right)}{\sum\limits_{i,j} K\left(\sqrt{(x')^2+(\frac{1}{\sigma_{x,y}}y')^2}\right)}, \tag{2}$$

where x' and y' are coordinates of points (i,j) in the coordinate system centered in the point (x,y) with the axe Ox directed along the gradient direction in the point (x,y):

$$x' = (x-i)\cos\theta_{x,y} + (y-j)\sin\theta_{x,y},$$
$$y' = -(x-i)\sin\theta_{x,y} + (y-j)\cos\theta_{x,y}.$$

The value $\theta_{x,y}$ is the angle of the gradient in the point (x,y) and $\sigma_{x,y}$ is the kernel deformation coefficient which depends on the value of the gradient $|g_{x,y}|$.

We use the following value

$$\sigma_{x,y} = \begin{cases} \sigma_0, & |g_{x,y}| \geq g_0, \\ 1 + (\sigma_0 - 1)\frac{|g_{x,y}|}{g_0}, & 0 \leq |g_{x,y}| < g_0, \end{cases}$$

where the values $\sigma_0 \geq 1$ and $g_0 > 0$ are the parameters of the proposed gradient method. Using $\sigma_0 = 1$ will convert the gradient image interpolation method (2) to linear interpolation method (1). These parameters are specific to the interpolated image class and the interpolation kernel $K(t)$.

For the bicubic interpolation kernel

$$K(t) = \begin{cases} (a+2)|t|^3 - (a+3)|t|^2 + 1 & \text{for } |t| \leq 1, \\ a|t|^3 - 5a|t|^2 + 8a|t| - 4a & \text{for } 1 < |t| < 2, \\ 0 & \text{otherwise.} \end{cases}$$

with $a = -0.5$ and images from LIVE image database [10,11] it was found that using the values $\sigma_0 = 1.5$ and $g_0 = 20$ maximizes the $SSIM$ metric value [13].

Fig. 1 illustrates the proposed gradient image interpolation method.

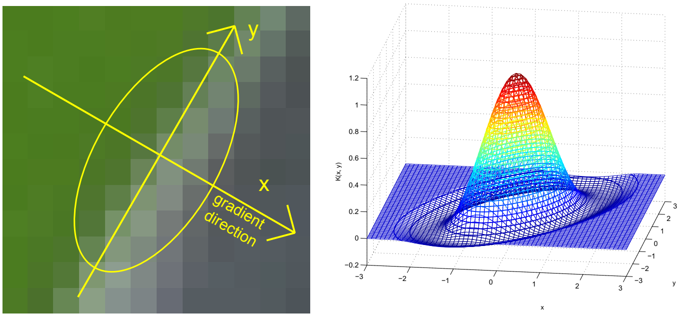


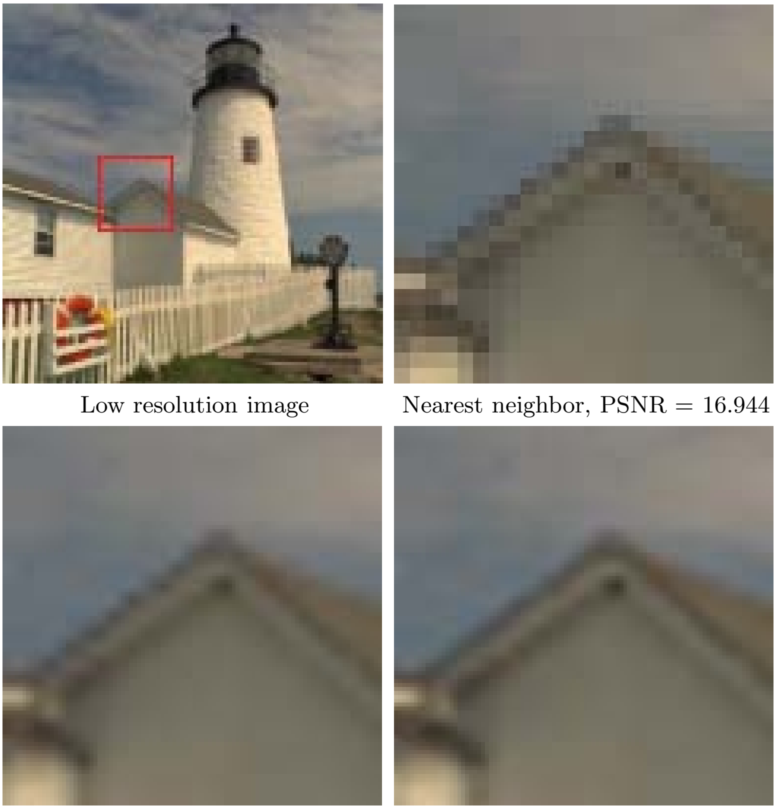
Fig. 1. The kernel function of the gradient based method in the point on the edge for $\sigma_{x_0,y_0} = 2$

The proposed gradient based image interpolation method shows better results than linear methods but worse than high quality non-linear algorithms like NEDI [7] or algorithms based on regularization [5]. Nevertheless the proposed method is significantly faster than non-linear image resampling algorithms.

3 Gradient Based Color Image Interpolation

The idea of color image interpolation is to use the multichannel image gradient proposed by Di Zenzo [3]. It finds the direction θ_0 that maximizes the rate of change function

$$F(\theta) = g_{xx}\cos^2\theta + 2g_{xy}\sin\theta\cos\theta + g_{yy}\sin^2\theta,$$



Low resolution image

Nearest neighbor, PSNR = 16.944

Bicubic interpolation,
PSNR = 18.630

Color gradient based interpolation,
PSNR = 18.775

Fig. 2. The results of gradient based 4 times image upsampling for color images with bicubic kernel function

where

$$g_{xx} = \left(\frac{\partial R}{\partial x}\right)^2 + \left(\frac{\partial G}{\partial x}\right)^2 + \left(\frac{\partial B}{\partial x}\right)^2,$$

$$g_{yy} = \left(\frac{\partial R}{\partial y}\right)^2 + \left(\frac{\partial G}{\partial y}\right)^2 + \left(\frac{\partial B}{\partial y}\right)^2,$$

$$g_{xy} = \frac{\partial R}{\partial x}\frac{\partial R}{\partial y} + \frac{\partial G}{\partial x}\frac{\partial G}{\partial y} + \frac{\partial B}{\partial x}\frac{\partial B}{\partial y}.$$

The explicit formula to find the direction of the multichannel gradient θ_0 looks as

$$\theta_0 = \frac{1}{2}\arctan\frac{2g_{xy}}{g_{xx} - g_{yy}}.$$

The value $F(\theta_0)$ is the gradient power.

We apply the grayscale gradient image interpolation method ([2]) to color images using Di Zenzo gradient. Fig. 2 shows the results of gradient image interpolation with the bicubic kernel function.

4 Improvement of Color Image Interpolation by the Gradient Based Method

The proposed gradient method cannot give results with quality compared to high-quality non-linear methods even when color image gradient is used. But processing of all components of the color image by high-quality methods is time consuming.

We suggest the algorithm for color image interpolation which takes a color low resolution image and a grayscale high resolution image obtained by a high-quality grayscale image interpolation method as input data. The grayscale high-resolution image is used to refine the color image gradient. The algorithm consists of the following steps:

1. Calculate the color high-resolution image with chrominance components interpolated from the given low-resolution image using bicubic interpolation and the luminance component taken from the given grayscale high-resolution image.

2. Perform interpolation of the given low-resolution image by the proposed color gradient based method with color gradient taken from the color high-resolution image obtained in the previous step.

3. Take chrominance components of the obtained in the previous step color image resolution image and add it to the given grayscale high resolution image. The obtained image is the result of the algorithm.

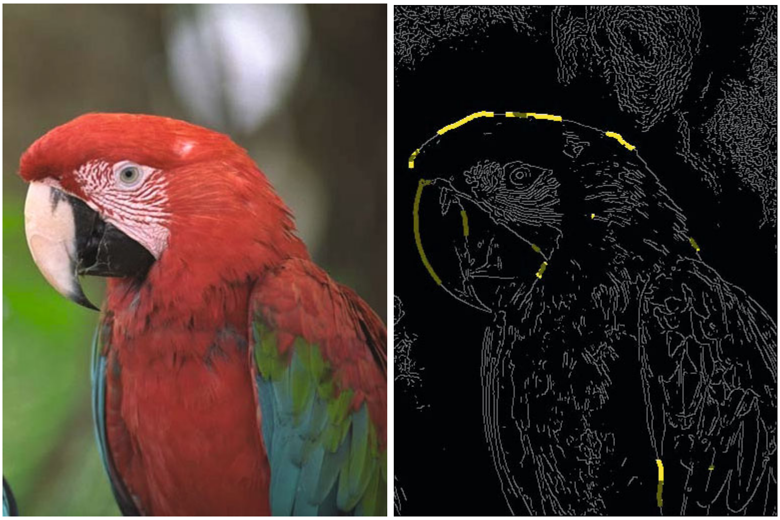
5 Color Basic Edges

To analyze the quality of the proposed algorithm, we seek for edges where the difference between gradient based interpolation and linear methods is noticeable.

For grayscale images, the concept of basic edges [9] is used to find the edges which can be used to estimate image quality. By the term 'basic edges' sharp edges distant from other edges are called. Blur and ringing effect are the most noticeable near these edges.

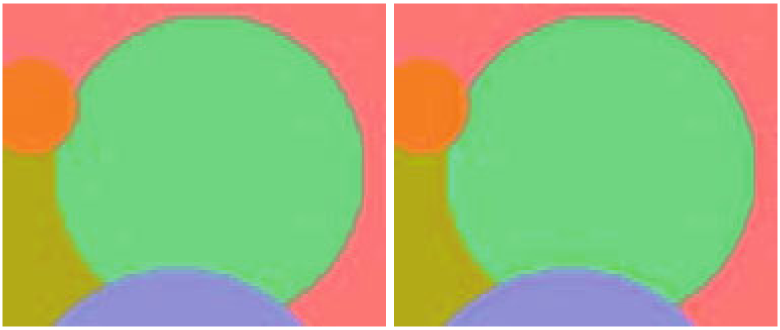
We extend the concept of basic edges to grayscale images using multichannel image gradient and color edge detection. But since human eye is more sensitive to intensity changes than to color changes, color artifacts are not visible if color basic edge is also a grayscale basic edge. Therefore, we seek for color basic edges which are not grayscale basic edges. We call these edges as pure color basic edges.

The example of color basic edges detection in shown in fig. 3.



The reference image. The result of basic edges detection.

Fig. 3. The results of finding basic edges areas for the image 'parrots'. Gray lines are image edges. Dark yellow area is the area of edges being both color and grayscale basic edges. Light yellow area is the area of color basic edges which are not detected as grayscale basic edges.



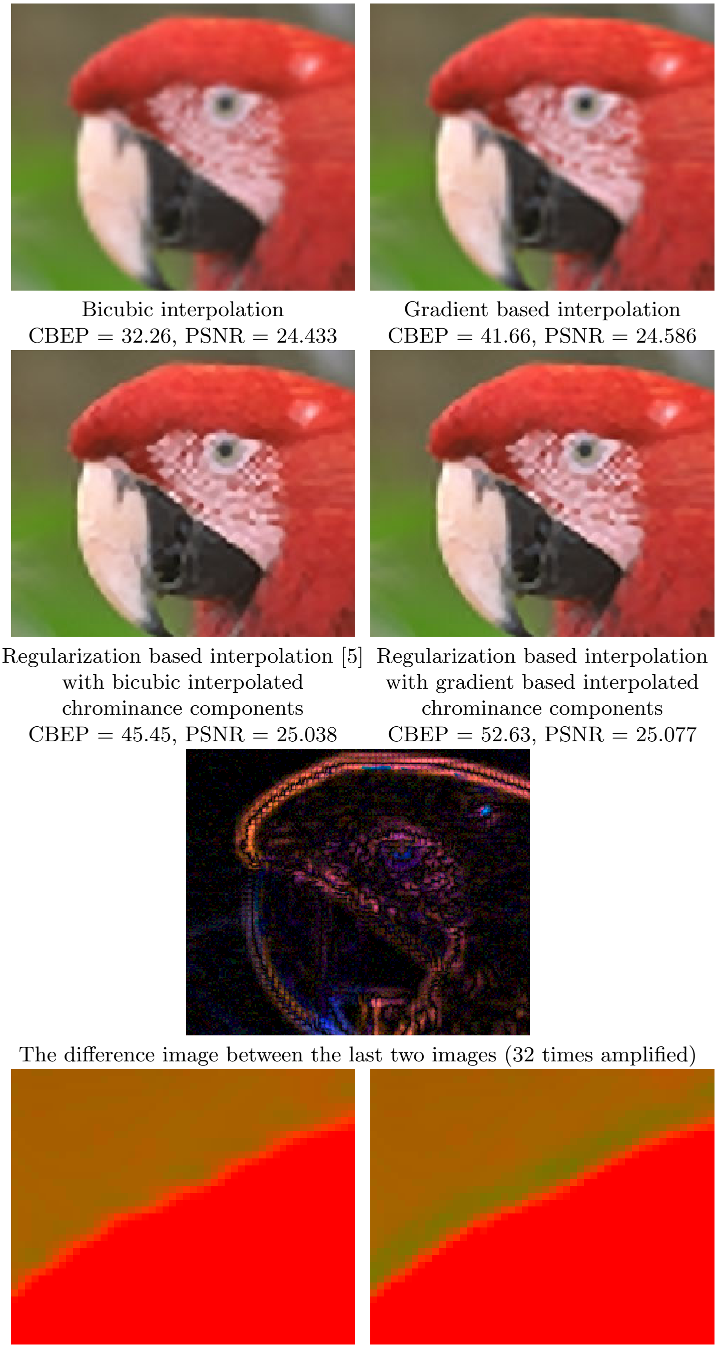
Bicubic interpolation
CBEP = 7.93, PSNR = 24.616

Gradient based interpolation
CBEP = 10.98, PSNR = 25.584

Fig. 4. The results of the proposed interpolation methods for the synthetic image with color edges only

6 Results

To evaluate the proposed method we took the reference images, downsampled it in 4 times, then applied regularization based image interpolation [5] and the proposed methods and compared the results with the reference images.



Bicubic interpolation
CBEP = 32.26, PSNR = 24.433

Gradient based interpolation
CBEP = 41.66, PSNR = 24.586

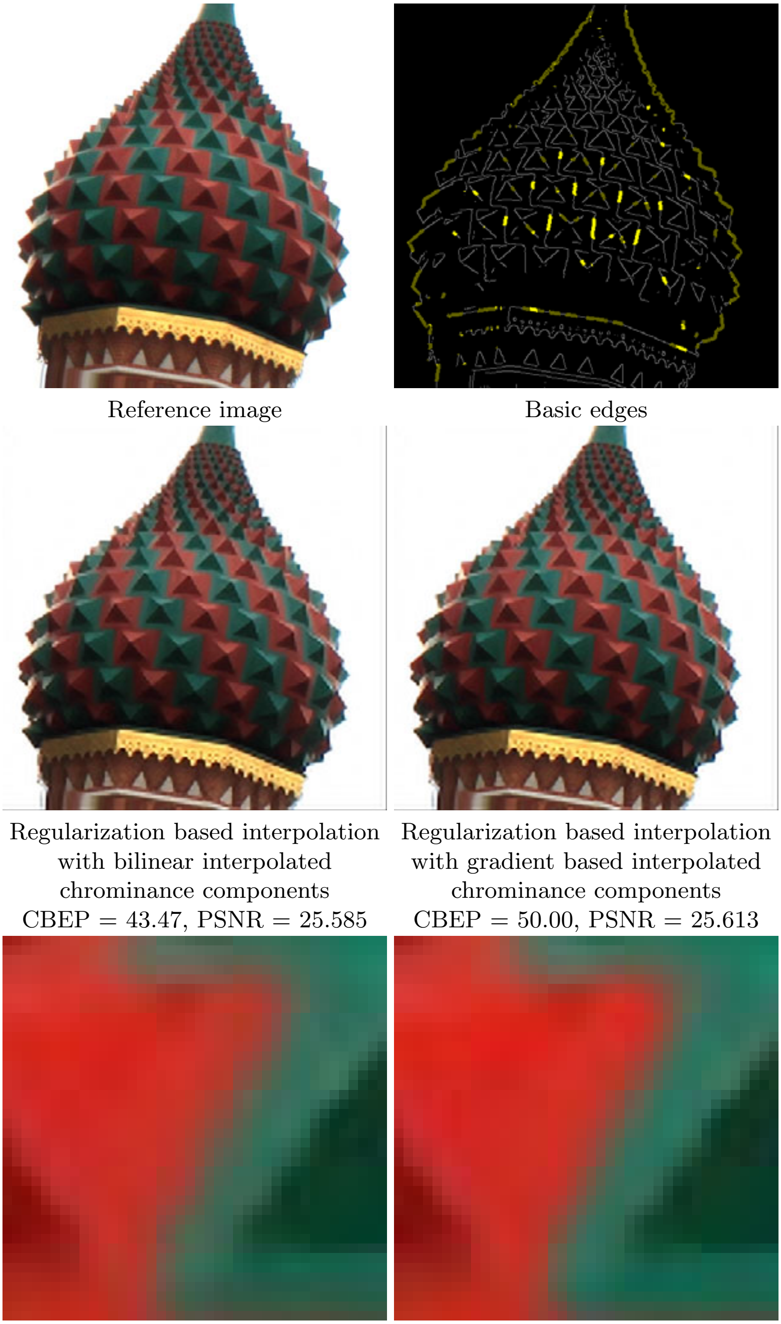
Regularization based interpolation [5]
with bicubic interpolated
chrominance components
CBEP = 45.45, PSNR = 25.038

Regularization based interpolation
with gradient based interpolated
chrominance components
CBEP = 52.63, PSNR = 25.077

The difference image between the last two images (32 times amplified)

Amplified image fragments of the last two images

Fig. 5. Application of the proposed color image interpolation methods to the image 'parrots'



Reference image

Basic edges

Regularization based interpolation with bilinear interpolated chrominance components
CBEP = 43.47, PSNR = 25.585

Regularization based interpolation with gradient based interpolated chrominance components
CBEP = 50.00, PSNR = 25.613

Amplified image fragments of the last two images

Fig. 6. Application of the proposed color image interpolation methods to the image 'cathedral'

Two metrics were calculated: PSNR and CBEP. The metric CBEP equals to DSSIM [13] metric calculated only in the area of interest M_{CBEP}. This area is the analog to BEP area introduced in [9] for color images. It consists of pixels in a small neighborhood of pure color basic edges. The metric PSNR was calculated for the entire image.

The results of the proposed interpolation methods for the synthetic image with only pure color basic edges are shown in Fig. 4. It can be seen that the gradient based method with bicubic kernel shows better edge quality than bicubic interpolation method.

For real images, the difference is less noticeable because of noise and few edges with completely absent grayscale gradient. In fig. 5 the results of the proposed color image interpolation methods for the image 'parrots' are shown. If bilinear interpolation is used for chrominance components, the difference is more visible. In fig. 6 the results for the image 'cathedral' with bilinear interpolation of color components are shown.

7 Conclusion

The gradient method for color image interpolation has been proposed. It uses the multichannel gradient to perform high quality interpolation of color edges that are not presented in the luminance image component.

The proposed algorithm enhances the grayscale high resolution image obtained by a non-linear image resampling method with bilinear or bicubic chrominance components resampling. The effect can be seen in the area of image color basic edges. This approach can be used to improve the performance of high quality but very slow image interpolation algorithms that process all components of the image. Instead of processing all components of the image, it is possible to resample only luminance component by high quality algorithm and to interpolate chrominance components by the proposed gradient method.

References

1. Blu, T., Thevenaz, P., Unser, M.: Linear interpolation revitalized. IEEE Trans. Image Proc. 13(5), 710–719 (2004)
2. Chena, M.J., Huanga, C.H., Leea, W.L.: A fast edge-oriented algorithm for image interpolation. Image and Vision Computing 23(9), 791–798 (2005)
3. Di Zenzo, S.: A note on the gradient of a multi-image. Comput. Vision Graph. Image Process. 33, 116–125 (1986)
4. Koschan, A.: A comparative study on color edge detection. In: Li, S., Teoh, E.-K., Mital, D., Wang, H. (eds.) ACCV 1995. LNCS, vol. 1035, pp. 574–578. Springer, Heidelberg (1996)
5. Krylov, A.S., Lukin, A.S., Nasonov, A.V.: Edge-preserving nonlinear iterative image resampling method. In: Proceedings of International Conference on Image Processing (ICIP 2009), pp. 385–388 (2009)
6. Lee, Y.J., Yoon, J.: Nonlinear image upsampling method based on radial basis function interpolation. IEEE Trans. Image Proc. 19(10), 2682–2692 (2010)

7. Leitao, J.A., Zhao, M., de Haan, G.: Content-adaptive video up-scaling for high-definition displays. In: Proceedings of Image and Video Communications and Processing 2003, vol. 5022, pp. 612–622 (2003)
8. Li, X., Gunturk, B., Zhang, L.: Image demosaicing: a systematic survey. Visual Communications and Image Processing 6822, 68221J–68221J–15 (2008)
9. Nasonov, A.V., Krylov, A.S.: Basic edges metrics for image deblurring. In: Proceedings of 10th Conference on Pattern Recognition and Image Analysis: New Information Technologies, vol. 1, pp. 243–246 (2010)
10. Sheikh, H., Sabir, M., Bovik, A.: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Trans. Image Proc. 15(11), 3440–3451 (2006)
11. Sheikh, H., Wang, Z., Cormack, L., Bovik, A.: Live image quality assessment database release 2, http://live.ece.utexas.edu/research/quality
12. Sun, J., Xu, Z., Shum, H.Y.: Image super-resolution using gradient profile prior. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2008), pp. 1–8 (2008)
13. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Proc. 13(4), 600–612 (2004)

Path Analysis in Multiple-Target Video Sequences

Brais Cancela, Marcos Ortega, Alba Fernández, and Manuel G. Penedo

Varpa Group, Department of Computer Science
University of A Coruña, Spain
{brais.cancela,mortega,alba.fernandez,mgpenedo}@udc.es

Abstract. Path analysis becomes a powerful tool when dealing with behavior analysis, i. e., detecting abnormal movements. In a multiple target scenario it is complicated to obtain each object path because of collision events, such as grouping and splitting targets, and occlusions, both total or partial. In this work, a method to obtain the similarity between different trajectories is presented, based in register techniques. In addition, an hierarchical architecture is used to obtain the corresponding paths of the objects in a scene, to cope with collision events. Experimental results show promising results in path analysis, enabling it to establish thresholds to abnormal path detection.

1 Introduction

In the security field, path analysis is a powerful tool for detection of abnormal behavior in a given scenario. An strange movement of an object could result as an abnormal behavior, which should be detected by a surveillance system. Typically, this kind of scenarios comprise series of individuals (or objects in general) crossing the scene simultaneously. In order to obtain a good approach for the path associated with each moving object, a robust target tracking must be implemented able of dealing with multiple objects. This involves target detection, classification and collision processing (grouping, splitting, leaving and entering the scene or partial occlusions). This is necessary because the system cannot make a mistake identifying each object. An error identifying a moving object could result in a wrong associated path, making the comparison useless.

In this work we present a methodology for path analysis under multiple-target condition, establishing an adequate framework for higher-level applications that require behavior analysis, such as video surveillance. This paper is organized as follows: section 2 discusses the state-of-the-art in this field and related work; section 3 describes our proposed multiple-target tracking method, whereas section 4 describes how to model each object path and the method used to determine the similarity between two of them; section 5 shows the method results and section 6 offers conclusions and future work.

2 Related Work

There are in the literature many different approaches for path analysis. One of the earliest methods uses a Self-Organizing Feature Map [1]. They encode every path with a

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 50–59, 2011.

fixed length, which is introduced into the SOM to classify the trajectory between normal and abnormal. It cannot cope with multiple targets, and needs to train the system with a priori knowledge. In [2], Hidden Markov Models are used to define each path. Orientation is also introduced. It track multiple objects around a scene, but it does not deal with collision events. A log-likelihood function is used to classify each event. It is also used by Rao et al. [3], who use a probabilistic model to define model trajectories, but it only detects one object in each scene. Each feature vectors of a trajectory is assumed to have Markovian dependence rather than being independent. Calderara et al. [4] use a mixture of Von Mises distributions created from an unsupervised training set. No more than one one is detected under this technique.

Our approach to this topic is based in the register techniques to determine path similarities. [5,6]. Each path is shown as a binary image, which is compared against others using restricted transformations to obtain the best fit. In addition, a multiple-target tracking based in an hierarchical architecture is performed in order to deal with the occlusion problem. None of the methods explained before can handle with this issue.

Our goal is to obtain an architecture which can enable us to to establish thresholds to abnormal path detection. This system must be robust against multiple-tracking main problems: grouping, splitting and total or partial occlusion events. Our methodology works as follows: first, a multiple-target tracking method is developed to distinguish every object in the scene. Second, a method to detect each object path is presented, based in the tracking method properties. Finally, each path is classified into normal or abnormal using a restricted register technique.

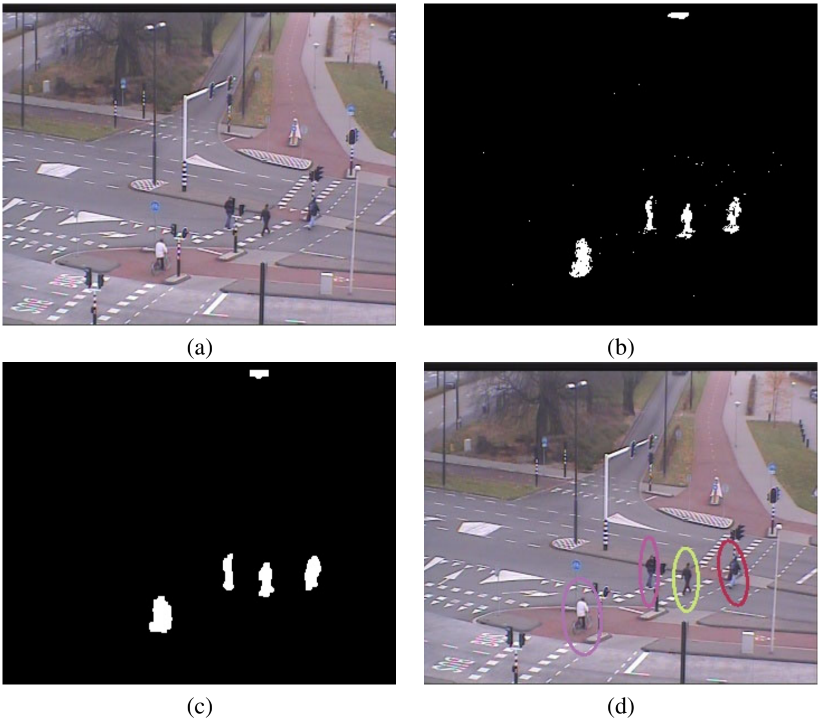
3 Multiple-Target Tracking

In order to cope with the problems of multiple-target tracking mentioned before, our system includes two different ways to detect an object: first, a low-level tracking is used to detect each object into the scene; second, a high-level representation must be implemented so that it could identify every moving object under different collision problems. Therefore, our method comprises a hierarchical architecture [7]. However, different changes are introduced to the original hierarchical model in order to further improve the accuracy.

First, a blob detection is performed to detect every moving object in the scene. A background subtraction method based in Mixture of Gaussians [8] (MoG) is used. Each background pixel is modeled as a bunch of five gaussians. Pixels which values do not fit under any background distribution are marked as foreground until there is a Gaussian distribution that includes them with enough evidence. Background model is updated with every new frame using a training parameter, which is the learning rate. To prevent the algorithm from stopped foreground pixels, which could be considered as background because of the updating system, the training parameter is set to a very low value. A requirement is that an only-background sequence is needed to train the algorithm.

After the background subtraction method is applied, the image is divided into foreground and background pixels. To fill the blobs and to avoid small regions holes or noise due to camera movement or video compression, opening and closing morphological operations and a minimum-area filter are applied when detecting blob regions.

Once detected, an ellipse representation for each blob is used, because it is really useful for dealing with collision problems, since it is a simple geometry which fits better with the object than polygons. Therefore, the j-observed blob at time t is given by $z_j^t = (x_j^t, y_j^t, h_j^t, w_j^t, \theta_j^t)$, where x_j^t, y_j^t represent the ellipse centroid, h_j^t, w_j^t are the major and minor axes, and θ_j^t the ellipse orientation. Fig. 1 shows an example of this methodology.


(a)

(b)

(c)

(d)

Fig. 1. (a) Frame. (b) MoG foreground detection. (c) Morphological operators. (d) Blob detection after applying minimum-area filter.

To track each moving object into successive frames, our approach stores the position of the ellipse in a window of size M. Subsequently, a median filter is used in order to smooth the values and a set of adaptive filters (Adalines) predict the velocity of each ellipse parameter. Adding this velocity to the previous position, a new predicted position is computed.

To match each ellipse with the appropriate tracking object, we look at the centroid point of the tracking object. Since we assumed that every object in the frame moves slowly enough compared to the frame rate, if the ellipse centroid is included into the ellipse of the new predicted position, and it is the only one, the tracking object is matched with the new ellipse and the low level tracking is confirmed. If two or more ellipse centroids are within the predicted position ellipse, then a splitting event is created. On the other hand, if two or more predicted position centroids are within a new ellipse, a grouping event is created.

To have the objects in the scene identified we propose an appearance model. This is needed because the low-level tracking has problems in cases of grouping and occlusion, since even although it can detect when a grouping or a splitting event occurs, it cannot detect which of the new blobs corresponds with a particular target stored into the group. As we mentioned before, a color representation of each target is computed. We select five different fixed color histograms, all of them into L*a*b space color. The histograms selected are the following:

$$\omega_1 * L + \omega_2 * a + \omega_3 * b, (\omega_1, \omega_2, \omega_3) \in \{(1,0,0); (0,1,0); (0,0,1); (0,1,1); (0,1,-1)\},$$

so that the histograms stored in each tracking object correspond with the appearance average of all objects along the sequence identified as the tracking object. This is needed because of light changes or orientation and position changes in the tracking object. With L*a*b color space we can isolate illumination into one component, which is useful to make this method invariant to fast illumination changes.

As we mentioned before, low-level tracking can detect grouping and splitting events, which are evaluated by the Event Management module. It defines six different states: single target, target grouping, grouped, splitting, split and occluded. When a grouping or splitting event occurs, if it is confirmed in the next frame, its state is changed to grouped or split, respectively. When, after t frames, typically the frame rate, a target does not appear, its state is changed to occluded. When it occurs, low-level tracking is not useful because the position is unpredictable after a long time.

When a low-level tracking is trained (detection after six consecutive frames), the Tracking Object Matching module is activated. After it is activated, every time the low-level tracker is confirmed, a feature comparison is computed to guarantee the low-level tracker confirmation is correct. If this module confirms the match, its state is updated with the new position and color histograms are updated. If no observation is associated to a particular target, its state is set using the previously predicted state.

When a target is marked as *occluded*, it means that low-level tracker is unused, so we only make an appearance comparison to locate the object. These trackers have lower priority than the others, meaning that they can only be compared when the rest of trackers failed.

If there is no matching between the low-level comparison in a long-duration occlusion, the tracker is marked as *occluded*. This means that we only make an appearance comparison to locate the object. However, these trackers have lower priority than the others, meaning that they can only be compared when the rest of trackers failed.

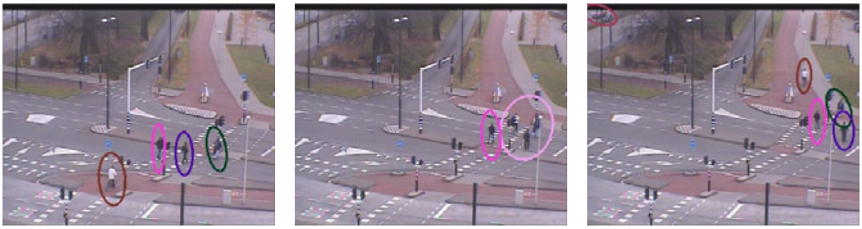
To make the appearance computation, we normalize and quantify each histogram into 64 bins to improve process velocity, which is high enough to prevent wrong matchings [7]. Thus, the probability of each feature is calculated as:

$$p_k^i = C^i \sum_{a=1}^{M} \delta(b(x_a) - k), \tag{1}$$

where C_i is a normalization constant which ensures $\sum_{k=1}^{64} p_k^i = 1$, δ is the Kronecker delta, $\{x_a; a = 1 : M\}$ represent the pixel locations, M is the number of target pixels, and $b(x_a)$ is a function that associates pixels to their corresponding bins.

Then, the similarity between two histograms is computed using the *Hellinger distance*, $d_H = \sum_{k=1}^{64} \sqrt{p_k q_k}$, where q_k are the bins stored in the tracked object. Similarity criterion acceptance uses both mean and standard deviation of the Hellinger Distance in all of the different matches, and they are updated every new match occurs. When at least 60% of the comparisons pass the test (three out of five histograms) we accept the match. If there is no match with any of the tracking objects, a new one is instantiated and trained. It does not mean that the new tracker could have never appeared before, so, once this is trained, it is compared against other *occluded* trackers, because it could be one of the previously defined. If this is the case, the tracker is merged with the previous one. Fig. 2 shows an example of a case in which it produces both grouping and splitting events. The method can keep correctly the identification of all the moving objects.

Finally, a tracker is deleted if it is lost before it is trained or if the number of times being present is much lower than the number of frames since it appeared for the first time.



(a) (b) (c)

Fig. 2. (a) Tracking before grouping. (b) Tracking during grouping. (c) Tracking after splitting event. The algorithm detects every object and can match all the objects in a correct way.

4 Path Analysis

As mentioned earlier, one of the most common uses for object tracking is the analysis of its behavior in the scene. This can be human actions, object interactions, etc. In this work we focus on path analysis to model object behavior. This can lead, for example, to the detection of suspicious trajectories in controlled environments. In our case, the detection of every object movement could define the most usual tracks around the scene.

Our model for path analysis comprises two step: a) Object path modeling and b) Path comparison methodology.

To model these trajectories, we use the properties defined in the Multiple-Target Tracking algorithm. As we mentioned before, this method needs to be able to distinguish each object among the others. Although the low-level tracker is unused when the moving object is lost, it will be used to obtain the path associated to each one. As each object is located within the scene by an ellipse, we only need to store its ellipse centroid in every frame to store its path on the scene. So the j-observed object path at time t is represented as

$$\mathcal{P}_j^t = \{(x_j^\tau, y_j^\tau) \in \mathbb{R}^2 \ \mid \ \tau = 0..t\} \tag{2}$$

Fig. 3 shows an example of different paths found in a scene. When a group of several objects is formed in a scene, a solution for their particular paths definition must be obtained. The intuitive notion of assigning the group ellipse centroid to all of the objects turns to be a bad idea when the ellipse is much greater than the individual ones as seen in Fig. 2-b. This is because the real location of the object is significantly displaced respect to the ellipse centroid. We propose a simple solution to overcome this problem in most of the situations. No positions are stored for particular objects while in a group. When the group is split, and some object is located as an individual again, the original path of the object is linked to the current one by linear interpolation. This solution is adequate assuming group path is coherent respect to the objects, which is true in the vast majority of cases.

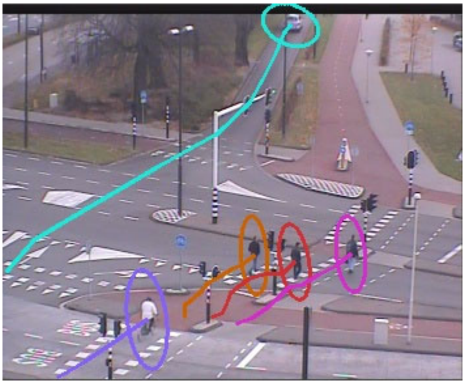


Fig. 3. Using the Multiple-Tracking framework, the path associated to each moving object involves the ellipse centroid, which represents the position over each time step.

Once the paths has been modeled, they can be compared and, therefore, a similarity measure must be defined to detect abnormal paths. In [9], an abnormal activity (abnormal trajectory in our case) is defined as a change in the system model, which could be slow or drastic, and whose parameters are unknown. In motion objects, there are a wide range of different situations that could be perceived as abnormal behavior, such as sudden changes in the speed or the direction. It could also be regarded as an abnormal situation when an object is stopped in its normal path [10].

In our case, our aim is to detect abnormal object movements with respect to the others, i. e., objects which path highly differs from the usual ones. Our first approach does not take into account neither orientation nor speed, but could be easily introduced in the model as this information is already known by the training system. Once the object has left the scene, its path is compared to determine if it may be regarded as an abnormal activity. Once we have all the points which determine the path of an object, a binary image is made, setting the path in white, as shown in Fig. 4.

To compare the path image against the others, we use a register technique, a method used in other fields such as biometrics, where templates are compared to each other to determine similarities. [5, 6]. The idea is to compare two templates, one used as reference and other that can be modified using geometric transformations in order to maximize the similarity by aligning the templates. In our case, the possible transformations is very limited and, therefore, we use a set of affine transformations allowing for small translations, scales and rotations. In order to measure the similarity between two aligned images, we use the normalized cross-correlation:

$$\mathcal{R}(x,y) = \frac{\sum_{x',y'}(T(x',y') \cdot I(x+x',y+y'))}{\sqrt{\sum_{x',y'} T(x',y')^2 \cdot \sum_{x',y'} I(x+x',y+y')^2}}, \tag{3}$$

where T is the reference template and I the one which can be modified.

5 Experimental Results

In our experiments we have used the CANDELA Intersection Scenarios [11] in order to test the methodology. Overall, we use over two minutes video recording at 25 frames per second. These videos take place outdoors. Partial occlusions, grouping and splitting events are frequent in these sequences making it a very suitable and challenging scenario to test our method.

To perform the abnormal activity detection, a few restrictions are inserted to guarantee a low rate of bad matches. First, we only evaluate $\mathcal{R}(0,0)$. This is because the nature of the path. For instance, a person who walks in the top of the image is not the same than other person walking in the bottom. This implies that we have to restrict the movement of the image against the template as much as it possible.

Second, several paths which are close together in a parallel way could be matched as the same track. To solve this situation, also improving the performance, we reduce the size of the image by a correction factor. This helps us to obtain more accurate results. To determine the correction factor value, we choose an exponential manner with powers of two. The value we use is the closest to the mean height of all the objects detected. Finally, restricted affine transformations are used. This includes rotations between -10 and 10 degrees and translations, both in x and y, between -1 and 1. A threshold is used to define if two path are the same or are different.

In Fig. 4 we can see an example of this methodology. Two different cars drive in the same road while one person is walking near them. Path images of both cars are close, while the other remains quite different. Using the restrictions explained before, we obtain a similarity value of 0.707776, while the similarity between the cars and the person is lower than 0.13, allowing to establish a confidence band to select a similarity threshold to consider two paths the same one in our domain.

Over the videos used to test this system, thirty one different paths were stored and compared. Fig. 5-a shows a bar plot containing cross-correlation values in two different classes, those which are the same path under our criteria and those which are different.

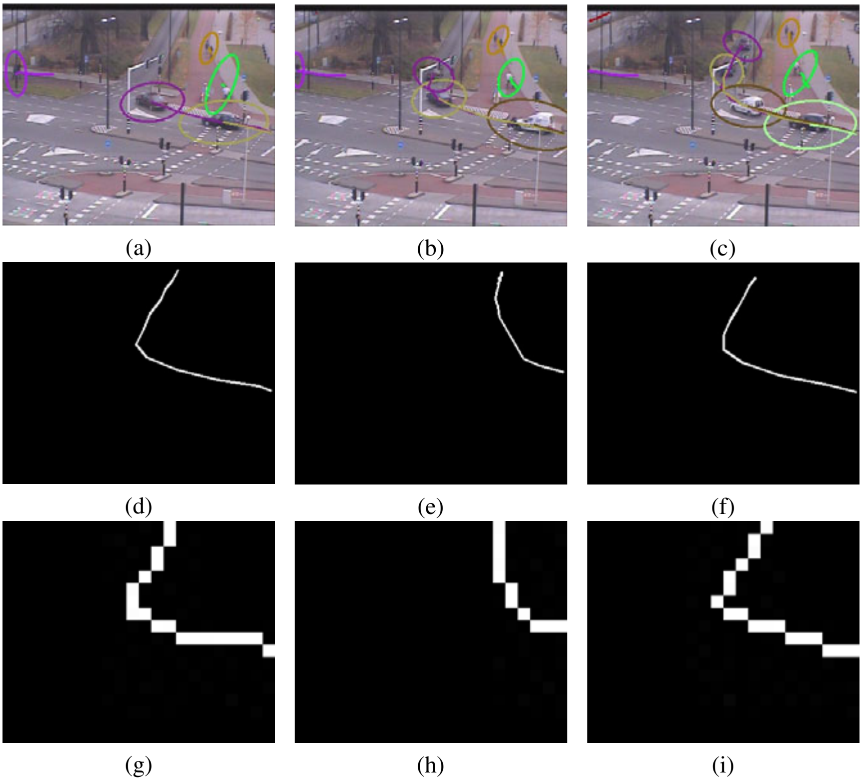


Fig. 4. Ellipse paths after a sequence. (a, b, c) frame sequence. (d) purple ellipse path template. (e) green ellipse path template. (f) ocher ellipse path template. (g) purple ellipse path template reduced by a correction factor. (h) green ellipse path template reduced by a correction factor. (i) ocher ellipse path template reduced by a correction factor.

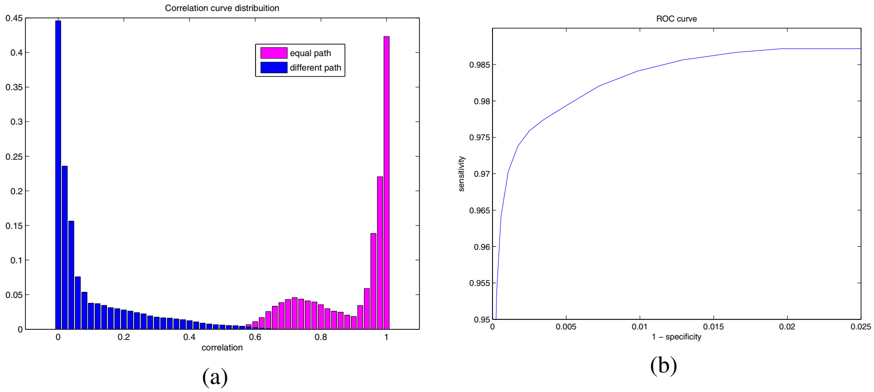


Fig. 5. (a) Correlation curve distribution. (b) ROC curve. Equal path and different path values are highly separated.

These distributions are highly separated, so we can obtain very good results. In Fig. 5-b we can see the ROC curve, which is close to the Heaviside step function. In fact, choosing the threshold value 0.56 we obtain the results showed in Table 1, which demonstrate the validity of the methodology for path analysis.

Table 1. Sensitivity and specificity of the method using the threshold value 0.56. Results show the accuracy of the methodology.

	Value
Sensitivity	98.75%
Specificity	98%

6 Conclusions

In this paper a new approach to the path detection for multiple object tracking is presented using an hierarchical architecture. Low-level tracker properties are used in order to obtain the path followed by every moving object detected in the scene. An abnormal path detection is used, based in register techniques with restrictions, obtaining promising results.

A multiple-tracking object is developed to operate under collision events, like grouping, splitting or total and partial occlusions. A register-based technique is used to obtain the similarity between trajectories, which allow us to establish thresholds to abnormal paths detection, obtaining over a 98.5% of success.

In a future research a dynamic methodology would be interesting to enable path comparisons even if the moving object have not left the scene. A methodology to detect different paths because of the different direction of the movement will be developed.

Acknowledgments. This paper has been partly funded by the Consellería de Industria. Xunta de Galicia through grant contracts 10/CSA918054PR and 10TIC009CT.

References

1. Owens, J., Hunter, A.: Application of the self-organising map to trajectory classification. In: Proceedings of Third IEEE International Workshop on Visual Surveillance (2000)
2. Jiang, F., Wu, Y., Katsaggelos, A.K.: Abnormal event detection from surveillance video by dynamic hierarchical clustering. In: IEEE International Conference on Image Processing, ICIP 2007, vol. 16 (2007)
3. Rao, S., Sastry, P.S.: Abnormal activity detection in video sequences using learnt probability densities. In: TENCON 2003, Conference on Convergent Technologies for Asia-Pacific Region, vol. 1, pp. 369–372 (2003)
4. Calderara, S., Cucchiara, R., Prati, A.: Detection of abnormal behaviors using a mixture of von mises distributions. In: IEEE Conference on Advanced Video and Signal Based Surveillance, AVSS 2007, pp. 141–146 (2007)

5. Mariño, C., Penedo, M.G., Penas, M., Carreira, M.J., González, F.: Personal authentication using digital retinal images. Pattern Analysis and Applications 9, 21–33 (2006)
6. Mariño, C., Ortego, M., Barreira, N., Penodo, M.G., Carreira, M.J., González, F.: Algorithm for registration of full scanning laser ophtalmoscope video sequences. Computer Methods and Programs in Biomedicine 102(1), 1–16 (2011)
7. Rowe, D., Reid, I., Gonzàlez, J., Villanueva, J.J.: Unconstrained multiple-people tracking. In: Franke, K., Müller, K.-R., Nickolay, B., Schäfer, R. (eds.) DAGM 2006. LNCS, vol. 4174, pp. 505–514. Springer, Heidelberg (2006)
8. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 246–252 (1999)
9. Vaswani, N., Chowdhury, A.R., Chellappa, R.: ”shape activity”: A continuous state hmm for moving/deforming shapes with application to abnormal activity detection. IEEE Transactions on Image Processing 14(10), 1603–1616 (2005)
10. Grimson, W., Lee, L., Romano, R., Stauffer, C.: Using adaptive tracking to classify and monitor activities in a site. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, Santa Barbara, CA, pp. 22–31 (1998)
11. CANDELA: content analysis and networked delivery architectures, `http://www.multitel.be/~va/candela/`

Statistical Multisensor Image Segmentation in Complex Wavelet Domains

Tao Wan[1] and Zengchang Qin[1,2]

[1] Robotics Institute, Carnegie Mellon University, Pittsburgh, USA,
[2] Intelligent Computing and Machine Learning Lab,
School of ASEE, Beihang University, Beijing, 100191, China
{taowan,zcqin}@andrew.cmu.edu

Abstract. We propose an automated image segmentation algorithm for segmenting multisensor images, in which the texture features are extracted based on the wavelet transform and modeled by generalized Gaussian distribution (GGD). First, the image is roughly segmented into textured and non-textured regions in the dual-tree complex wavelet transform (DT-CWT) domain. A multiscale segmentation is then applied to the resulting regions according to the local texture characteristics. Finally, a novel statistical region merging algorithm is introduced by measuring a Kullback-Leibler distance (KLD) between estimated GGD models for the neighboring segments. Experiments demonstrate that our algorithm achieves superior segmentation results.

Keywords: multisensor image segmentation, statistical modeling, complex wavelets.

1 Introduction

The purpose of image segmentation is to produce a partition of the underline scene so that there are similar intensity or texture characteristics within each partitioned region. These features would favour a multisensor image fusion algorithm performed on each region in the fusion process. For this specified application, the segmentation should ideally have the following properties:

- All the salient objects, especially small-sized, in the input images are detected by the segmentation algorithm;
- All required features should be segmented as single separate regions;
- The segmented images should be able to provide precise region boundaries.

Although the above criteria can also be applied to a general task for segmenting natural images, they are the key factors of obtaining effective segmentations from the available sensors to perform a successful region-based image fusion (Lewis et al. 2007, Wan et al. 2009). In this paper, we present an automated image segmentation algorithm to fulfill such goals.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 60–68, 2011.

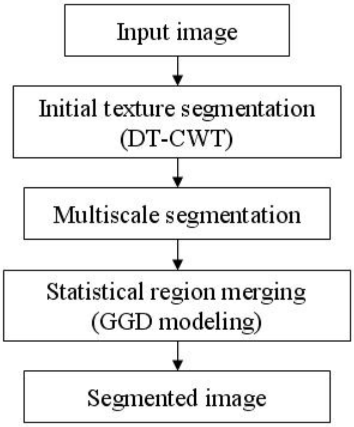

Input image
Initial texture segmentation (DT-CWT)
Multiscale segmentation
Statistical region merging (GGD modeling)
Segmented image

Fig. 1. The proposed image segmentation algorithm

In the previous work (Wan et al. Apr. 2007), we have developed a multiscale image segmentation algorithm based on the dominant color and homogeneous texture features. A statistical approach (Wan et al. Sep. 2007) using a non-Gaussian model was also proposed for segmenting natural color images. This paper introduces alternative approaches to the texture feature extraction and the region merging components of the algorithm in (Wan et al. Apr. 2007, Wan et al. Sep. 2007) to process multisensor images. The texture feature extraction is improved by performing in the wavelet domain, while the region merging is based on the Kullback-Leibler distance (KLD) (Do and Vetterli 2002) as a similarity metric between two neighboring segments.

As shown in Fig.1, the segmentation algorithm comprises three components. First, the image is roughly segmented into textured and non-textured regions using the dual-tree complex wavelet transform (DT-CWT) (Kingsbury 2001). The initial texture map is generated to discover the small salient objects that may be neglected by the following segmentation procedure. A multiscale segmentation is then applied to the resulting regions by considering different local texture features. A boundary refinement step is utilized to improve the segmentation performance. Finally, a statistical region merging approach is developed to measure the KLD between two generalized Gaussian models corresponding to the adjacent segments, ensuring the important features are grouped into individual regions. This three-component method is specifically designed to meet the criteria that could be useful in the segmentation-driven image fusion application. The key to the success of the proposed algorithm is that the statistical modeling techniques and wavelet textures are integrated into an unified framework to achieve precise and robust segmentation for multisensor images.

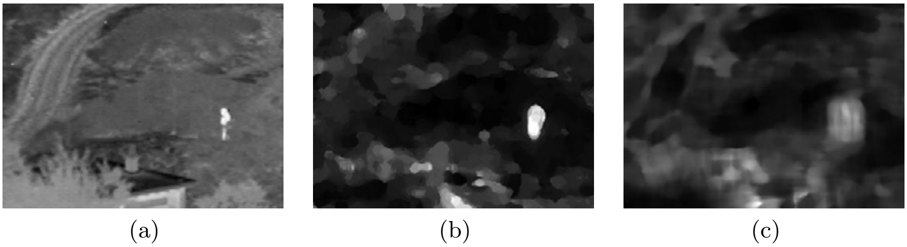

(a) (b) (c)

Fig. 2. The resulting images by using the truncated median filter. (a) "UNCamp" IR image. (b) The filtered image using DT-CWT. (c) The filtered image using Gabor decomposition.

2 Initial Texture Segmentation

Wavelets have emerged as an effective tool to analyze texture information as they provide a natural partition of the image spectrum into multiscale and oriented subbands. In this work, we use a three-scale DT-CWT (Kingsbury 2001) with six orientations, which is able to provide approximate shift invariance and directional selectivity while preserving the usual properties of perfect reconstruction and computational efficiency. The feature value $T(x, y)$ at the pixel location (x, y) is defined as:

$$T(x,y) = \{w_{l,\theta}(x,y)\} \qquad l = 1, 2, 3 \quad \theta = 1, 2, \ldots, 6 \tag{1}$$

where $w_{l,\theta}$ are the DT-CWT coefficients in the l^{th} level and θ^{th} orientation band.

A truncated median filter (Nixon and Aguado 2008) is applied to $T(x, y)$ of each subband to filter out the texture associated with transitions between regions. Fig.2 shows the filtered images using the DT-CWT and Gabor decomposition on an infrared (IR) image. By examining the figure, we can see that the DT-CWT produces more desirable results by highlighting the salient objects as well as maintaining the clear and smooth boundaries, which leads to a better quality image texture map.

A two-level K-means algorithm is used to define textured and non-textured regions. A pixel is classified as textured if the proportion of the number of the subbands belonging to the textured region is above a threshold P. Our experiments show that an appropriate value for thresholding the multisensor images is assigned to $P = 0.7$. The final texture maps are illustrated in Fig.3, in which the DT-CWT demonstrates a strong capability to detect the small-sized salient objects.

3 Multiscale Image Segmentaion

The textured and non-textured regions are further segmented into relatively small and homogeneous regions while retaining the boundaries between the two

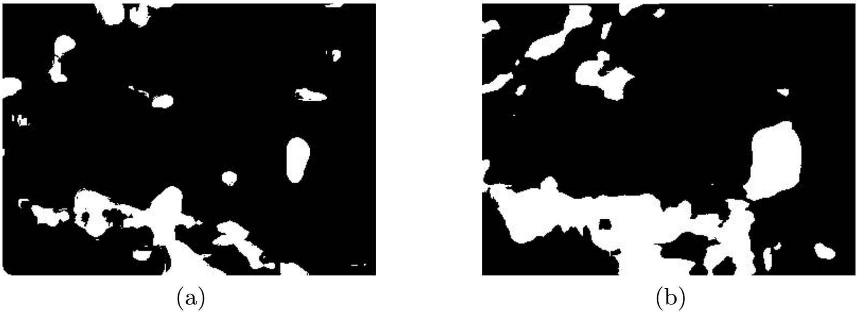


Fig. 3. The texture maps of "UNCamp" IR image. (a) The texture map using DT-CWT. (b) The texture map using Gabor decomposition.

regions. The dominant grayscale values are first extracted based on the peer group filtering and generalized Lloyd algorithm (Deng and Manjunath 2001). Then, the *JSEG* algorithm proposed in (Deng and Manjunath 2001) is used to minimize the cost associated with partitioning an image at different scales. A bigger window size is used for high scales, which are useful for detecting texture boundaries, while lower scales are employed in order to localize the intensity of grayscale edges. It is reasonable to apply the lower scales to the non-textured region, which has a more or less homogeneous texture, while higher scales are adopted for the textured region to find the texture boundaries. In contrast with the *JSEG*, which does not account for the local texture difference between the image regions, the strength of this approach is that we are able to apply the multiscale segmentation simultaneously to the same image according to the local texture characteristics.

Nevertheless, the resulting boundary locations between textured and non-textured regions are not the actual boundaries due to the fact that K-means clustering can only segment the image into rough regions. Moreover, multiscale segmentation provides accurate results only within the textured and non-textured regions. Consequently, a boundary refinement step is employed to adjust the boundaries between the two regions. A pixel is assigned to the neighbor class that has the minimum D value using the following function:

$$D = Dist(G^0, G^j) + a(S_4^j - D_4^j) + b(S_8^j - D_8^j) \tag{2}$$

where $Dist$ refers to the Euclidean distance measure, G^0 and G^j are the grayscale intensity of the current pixel and its j^{th} neighboring segment, S_4^j and S_8^j are the numbers of 4 and 8-neighbor pixels belonging to the j^{th} segment, while D_4^j and D_8^j are the numbers of 4 and 8-neighbor pixels belonging to the different classes of the j^{th} segment. a and b represent the strength of the spatial constraint. Specifically, as a and b increase, a pixel is more likely to belong to the class to which many of its neighbors belong. Thus region boundary smoothness is achieved. In all the experiments, a and b are assigned values of 0.8 and 0.6.

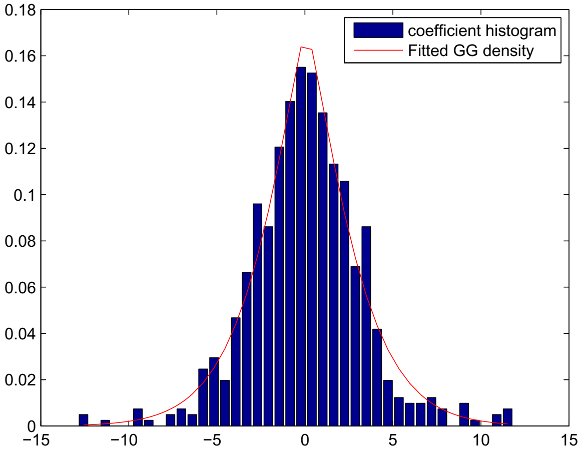


Fig. 4. An example of a region of the 1^{th} level and 2^{th} orientation DT-CWT subband coefficients histogram fitted with a generalized Gaussian density on a log scale. The estimated parameters are: $\alpha = 3.1370$, $\beta = 1.2782$.

4 Statistical Region Merging

After applying a multiscale segmentation on textured and non-textured regions, there are usually some very small regions or more than one neighboring regions exhibiting similar attributes. A merging step is therefore used to eliminate such instances by generating bigger regions. Do and Vetterli recently introduced a statistical framework for texture retrieval in content-based image retrieval applications (Do and Vetterli 2002), where the subband marginal density of the coefficients was approximated by the generalized Gaussian distribution (GGD), and KLD was computed as a similarity measurement. In this study, a statistical method is implemented by varying two model parameters α and β of the GGD to appropriately model the wavelet coefficients within the segmented regions. The segments with more than 80% of their pixels belonging to the non-textured area are categorized as non-textured segments, and the remaining segments are classified as textured segments. Therefore, segmented regions are considered individually rather than globally.

A corresponding merging criterion is provided for each category. The difference lies in the way the features are extracted within the regions. Non-textured segments are merged based on their grayscale intensity similarity. To achieve this, the Euclidean distance of the gray-level histograms extracted from the neighboring non-textured segments is calculated. For textured segments, region similarity is measured using the statistical model parameters followed by computing the Kullback-Leibler distance. We model the wavelet coefficients of each textured

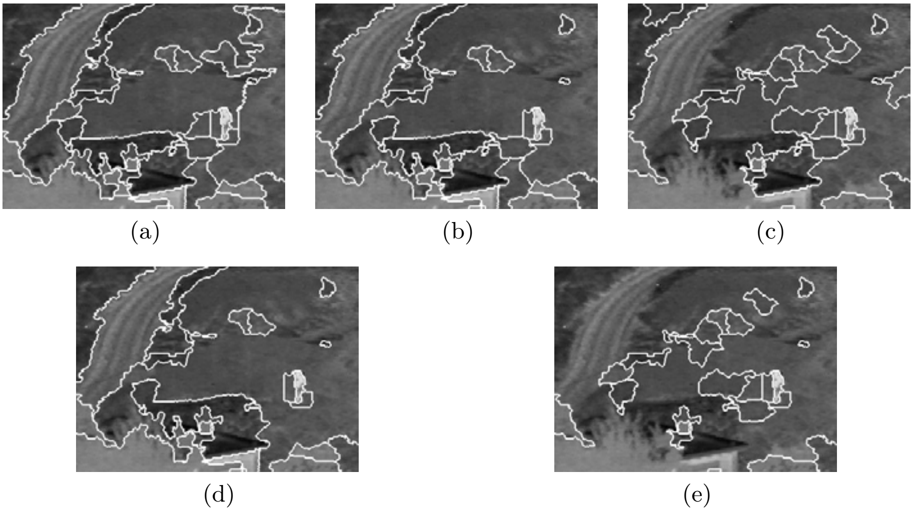

(a) (b) (c) (d) (e)

Fig. 5. The region merging results for "UNCamp" IR image. (a) The input segmented image. (b) The segmented image using KLD with 31 segments. (c) The segmented image using Euclidean distance with 31 segments. (d) The segmented image using KLD with 26 segments. (e) The segmented image using Euclidean distance with 26 segments.

region independently by a generalized Gaussian distribution. The KLD between two adjacent textured segments is defined as:

$$KLD(s_1, s_2) = \frac{1}{18}\sum_{l=1}^{3}\sum_{\theta=1}^{6}\log\left(\frac{\beta_1^{l,\theta}\alpha_2^{l,\theta}\Gamma(1/\beta_2^{l,\theta})}{\beta_2^{l,\theta}\alpha_1^{l,\theta}\Gamma(1/\beta_1^{l,\theta})}\right) + \left(\frac{\alpha_1^{l,\theta}}{\alpha_2^{l,\theta}}\right)^{\beta_2^{l,\theta}} \frac{\Gamma((\beta_2^{l,\theta}+1)/\beta_1^{l,\theta})}{\Gamma(1/\beta_1^{l,\theta})} - \frac{1}{\beta_1^{l,\theta}} \tag{3}$$

where $\Gamma(\cdot)$ is the Gamma function, s_1 and s_2 are the adjoining texture segments, and l and θ denote the index of the decomposition level and orientation. The characteristics of the regions can be completely defined via two model parameters α and β of the GGD. α and β can be estimated using the second (m_2) and fourth (m_4) order moments of image coefficients (Simoncelli and Anderson 1996).

$$m_2 = \frac{\alpha^2\Gamma(\frac{3}{\beta})}{\Gamma(\frac{1}{\beta})} \qquad m_4 = \frac{\Gamma(\frac{1}{\beta})\Gamma(\frac{5}{\beta})}{\Gamma^2(\frac{3}{\beta})} \tag{4}$$

Fig.4 demonstrates a typical example of a histogram of detail coefficients in a particular subband within a region together with a plot of the fitted estimated GGD model. The fits are generally good. As a result, with only two parameters of the GGD, we can accurately capture the characteristics of marginal densities of wavelet subband coefficients in each region.

Fig.5 shows the mediate merging results for “UNCamp” IR image after few iterations of the region merging algorithm using the KLD and Euclidean distance. Each pair of images for the comparison contain the same number of regions. The figure displays the actual merging process as similar neighboring regions are integrated into one segmented region. It is clear that the KLD provides better results than the Euclidean distance in terms of human perception.

5 Experimental Results and Discussions

The segmentation algorithm has been subjectively evaluated on various images of the same scene captured by different sensors. In Fig.6 and Fig.7, we demonstrate two examples of segmented results in comparison with the *JSEG* (Deng and Manjunath 2001) and the *Watershed* (O’Callaghan and Bull 2005). The *JSEG* examples shown are optimized using the best option values through visual inspection. The *Watershed* applies the region-depth threshold to the gradient surface which is set as 0.15 times the median gradient. Two pairs of images are used in the experiments, including the IR and visible images, a magnetic resonance image (MRI) and a computer tomography (CT) image. For the first pair. the hand-labeled ground truth segmentation are available from the multisensor image segmentation dataset (Lewis et al. 2006). In Fig.6(a), the important regions, such as road, house, and human figure shown in the ground truth, are well segmented by the proposed method, but over-segmented in the *JSEG*, and half merged with the scene background in the *Watershed*. This also happens to the visible image. For the CT image, there are no distinct differences between our method and the *Watershed*. However, in the *JSEG* segmentation shown in Fig.7(b), the central part of the structure has not been segmented and merged into the background. In Fig.7(d), both inner soft tissues and outer structure in

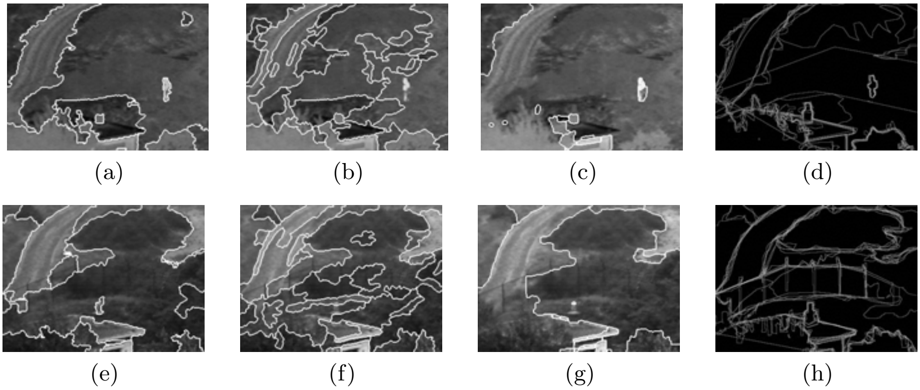


Fig. 6. The segmentation results for “UNCamp” IR and visible images. From left to right: the proposed segmentation algorithm, *JSEG* , *Watershed*, hand-labeled ground truth segmentation.

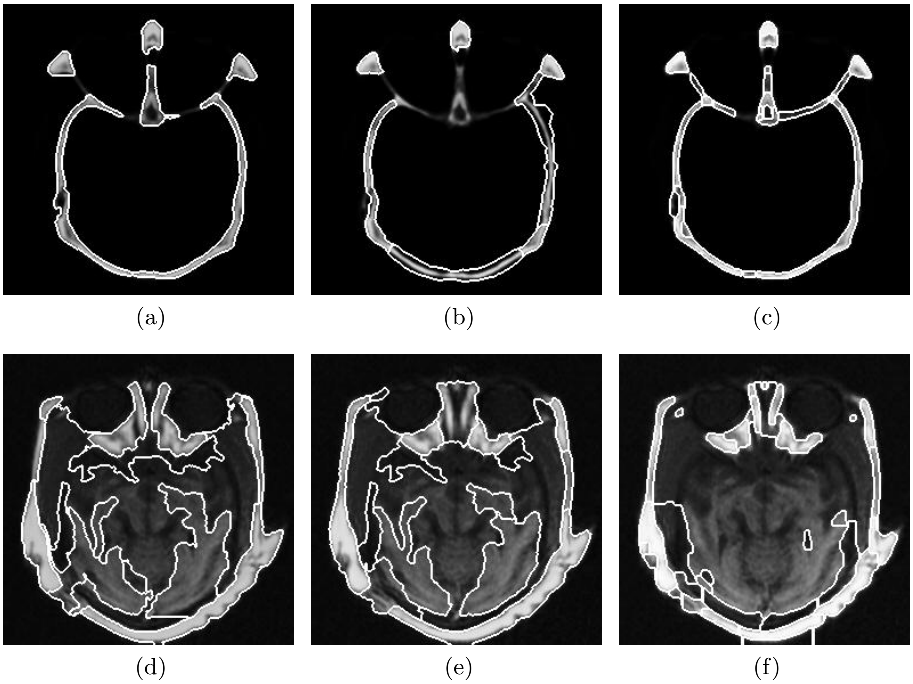


Fig. 7. The segmentation results for CT and MRI images. From left to right: the proposed segmentation algorithm, *JSEG*, *Watershed*.

the MRI are better segmented than the *JSEG* and *Watershed*. The experiments demonstrate that the proposed method provides an effective may to accurately segmenting the meaningful objects, even in small size, into separate regions.

6 Conclusions and Future Work

We have demonstrated that robust and meaningful image segmentation can be achieved by integrating the statistical modeling with the wavelet feature extraction and multiscale segmentation. The proposed algorithm is evaluated on a variety of multisensor images. The results show the effectiveness of the segmentation for multisensor images. In future, we are interested in applying the segmentation algorithm to a region-based image fusion scheme.

Acknowledgment. This work is partially funded by the NCET Program of MOE, China, the SRF for ROCS and the China Scholar Council for visiting fellowship (No. 2010307502).

References

Deng, Y., Manjunath, B.: Unsupervised Segmentation of Color-Texture Regions in Image and Video. IEEE Tran. Pattern Anal. Machine Intell. 23, 800–810 (2001)

Do, M., Vetterli, M.: Wavelet-Based Texture Retrieval Using Generalized Gaussian Density and Kullback-Leibler Distance. IEEE Tran. Image Process. 11, 146–158 (2002)

Kingsbury, N.: Complex Wavelets for Shift Invariant Analysis and Filtering of Signals. Appl. Compt. Harmon. Anal. 10, 234–253 (2001)

Lewis, J., Nikolov, S., Canagarajah, N., Bull, D., Toet, A.: Uni-Modal Versus Joint Segmentation for Region-Based Image Fusion. In: Proc. of the Int. Conf. Information Fusion, pp. 1–8 (2006)

Lewis, J., O'Callaghan, R., Nikolov, S., Bull, D., Canagarajah, N.: Pixel- and Region-Based Image Fusion with Complex Wavelets. Information Fusion 8, 119–130 (2007)

Nixon, M., Aguado, A.: Feature Extraction and Image Processing. Academic Press, Oxford (2008)

O'Callaghan, R., Bull, D.: Combined Morphological-Spectral Unsupervised Image Segmentation. IEEE Tran. Image process. 14, 49–62 (2005)

Simoncelli, E., Anderson, E.: Noise Removal via Bayesian Wavelet Coring. In: Proc. of the IEEE Int. Conf. Image Process., pp. 378–382 (1996)

Wan, T., Canagarajah, N., Achim, A.: Multiscale Color-Texture Image Segmentation with Adaptive Region Merging. In: Proc. of the IEEE Int. Conf. Acoustics, Speech, and Signal Process., pp. 1213–1216 (2007)

Wan, T., Canagarajah, N., Achim, A.: Statistical Multiscale Image Segmentation via Alpha-Stable Modeling. In: Proc. of the IEEE Int. Conf. Image Process., pp. 357–360 (2007)

Wan, T., Canagarajah, N., Achim, A.: Segmentation-Driven Image Fusion Based on Alpha-Stable Modeling of Wavelet Coefficients. IEEE Tran. Multimedia 11, 624–633 (2009)

Activity Discovery Using Compressed Suffix Trees

Prithwijit Guha[1], Amitabha Mukerjee[2], and K.S. Venkatesh[2]

[1] TCS Innovation Labs, New Delhi
prithwijit.guha@tcs.com
[2] Indian Institute of Technology, Kanpur,
{amit,venkats}@iitk.ac.in

Abstract. The area of unsupervised activity categorization in computer vision is much less explored compared to the general practice of supervised learning of activity patterns. Recent works in the lines of activity "discovery" have proposed the use of probabilistic suffix trees (PST) and its variants which learn the activity models from temporally ordered sequences of object states. Such sequences often contain lots of object-state self-transitions resulting in a large number of PST nodes in the learned activity models. We propose an alternative method of mining these sequences by avoiding to learn the self-transitions while maintaining the useful statistical properties of the sequences thereby forming a "compressed suffix tree" (CST). We show that, on arbitrary sequences with significant self-transitions, the CST achieves a much lesser size as compared to the polynomial growth of the PST. We further propose a distance metric between the CSTs using which, the learned activity models are categorized using hierarchical agglomerative clustering. CSTs learned from object trajectories extracted from two data sets are clustered for experimental verification of activity discovery.

1 Introduction

Activities are the manifestations of transitions of spatio-temporal relations of the scene objects among themselves and/or the scene (Figure 1); and are modeled as sequences of object appearance/motion features – e.g. human pose sequences, vehicle trajectories etc. Such features are often selected using domain specific priors (e.g. human body models using articulated cylinders/ellipsoids, stick models etc. for gesture recognition [1]) or constructed directly from object features obtained from image data (e.g. human contour descriptors using Point Distribution Models [2]).

The selection of these object states dictate the classes of the activities that they might describe. However, we further require efficient statistical sequence modeling techniques for representing significant temporal patterns from the time-series data of the object states. Existing literature on activity analysis is vast and is mostly based on the supervised framework where, the hidden Markov models [3] and its variants (e.g. parameterized HMM [4], coupled HMM [5] etc.)

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 69–78, 2011.

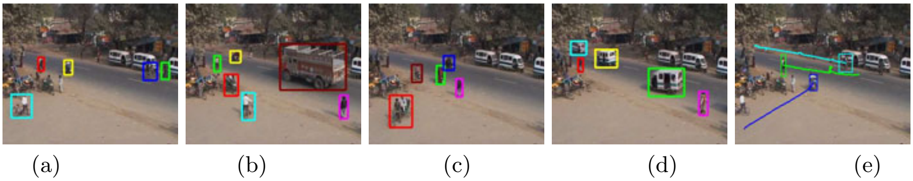


Fig. 1. Activity discovery – (a)-(d) Multiple object tracking performed on traffic video to extract (e) object trajectories which form important descriptors of object actions. These trajectories are modeled as *compressed suffix trees* (Section 2) which are clustered (Section 3) further to discover the action categories.

have been widely used for activity analysis. In comparison, much less has been explored in the domain of unsupervised activity modeling and recognition. An early work in this direction was proposed through categorizing object actions by trajectory clustering [6]. The use of variable length Markov models (VLMM) in the domain of activity analysis was introduced in [2] for modeling interactions. These approaches propose to perform a vector quantization over the object feature space to generate temporally indexed object-state sequences from video data. These sequences are parsed further to learn probabilistic suffix trees (PST) leading to the discovery of behavioral models of varying temporal durations. A good overview on contributions in such prior free activity modeling can be found in [7]. In the similar lines, activities have been modeled by stochastic context free grammars (SCFG) from sequences of scene landmarks traversed by the object [8]. These approaches have so far focused only on the activity modeling using VLMM, SCFG or PST. The issue of clustering such models have been addressed in [9] where activities were modeled from object state sequences in terms of PST which were further categorized through a graph based clustering using a similarity measure between the PSTs.

Our work develops along the lines of the approaches of "discovery" rather than supervised learning from manually labeled exemplars. We observe that the time indexed object descriptor sequences used for activity modeling are mostly populated with self-transitions. We argue that learning such self-transitions into probability suffix trees has minimal usage in the context of activity structure discovery and hence propose to learn "compressed suffix trees" (CST, Section 2). We further propose a semi-metric measure of dissimilarity between two CSTs (Section 3) and the *activity discovery* is performed by grouping the learned models using average linkage hierarchical agglomerative clustering. The experimental results on 2 data sets are presented for object actions modeled as sequences of quantized motion directions (Section 4).

2 Compressed Suffix Tree

Activity analysis applications have traditionally used the Hidden Markov Model (HMM) or its variants for classifying temporal patterns of action descriptors.

However, from the viewpoint of activity "*discovery*", this approach has two significant drawbacks – first, the supervised learning framework of HMM limits it from discovering the temporal pattern structures of the activities and second, the first order Markovian assumption adopted in the HMM fails to capture the essence of variable length constituent sub-events of the activities. This led the researchers to use variants of sequence mining algorithms such as the variable length Markov model [2], stochastic context free grammar [8] and probabilistic suffix tree [9] which showed satisfactory performance in discovering variable length structures of activities.

Given an input symbol sequence like "a a a b b c", a sequence mining algorithm aims at finding all sub-sequences of varying length and frequency – say, "a" occurred thrice, "b" twice, "a a" twice, "a b" once etc. When applied to activity discovery, these symbols are nothing but the object states - e.g. quantized motion directions of traffic objects, human body poses, spatio-temporal relations between interacting objects etc.

We observe that, the sequence mining algorithms encode the self-transitions as well – i.e. the self-transition $a \rightarrow a$ is also learned along with $a \rightarrow b$. In most practical cases, however, the same state may continue for a long time in such activity descriptor sequences. For example, consider the action of a car turning – moving westwards in the scene for 140 frames, north-westwards for 10 frames and northwards for the next 100 frames. Mining this motion direction sequence will result to the discovery of a lot of variable length sub-sequences like "westwards-westwards", "northwards-northwards-northwards" etc.

Object state transitions plays a major role in defining the temporal structure of an activity. Thus, we believe that learning only the transitions between different states is more useful compared to the maintenance of information regarding the self transitions. The direct advantage of avoiding to learn the self-transitions is reflected in the memory saving as the sub-sequences like $a \rightarrow a \rightarrow a$, $b \rightarrow b$ etc. need not be stored any more. This involves the removal of the redundant symbol repetitions from the original sequence by editing it to a "*self-transitionless*" sequence. Thus, We propose to perform sequence mining by avoiding self-transitions but preserving the relative frequencies of the symbols themselves and their transition sub-sequences. A formal description of the proposed approach of learning the "compressed suffix tree" (CST) and its operational dissimilarity from a sequence mining algorithm (say, PST) is presented next.

Consider a temporally ordered symbol sequence $\mathbf{S} = \{\epsilon(t), t = 1, \ldots\}$ where the symbols belong to some alphabet $\mathcal{E}$. The results of sequence mining are represented by a tree $\mathcal{T}$ rooted at a node ρ, say. Each node of $\mathcal{T}$ is a 2-tuple $\mathbf{T}_n \equiv (\epsilon, \pi)$ containing the symbol $\epsilon \in \mathcal{E}$ and a real number $\pi \in (0, 1]$ signifying the probability of occurrence of the path $\{\rho, \ldots \mathbf{T}_n\}$ among the set of all possible paths of the same length. However, the sequence mining is performed to discover variable length sequences only up to a certain maximum size L to maintain finite size of the tree. In other words, the maximum depth of the tree is L.

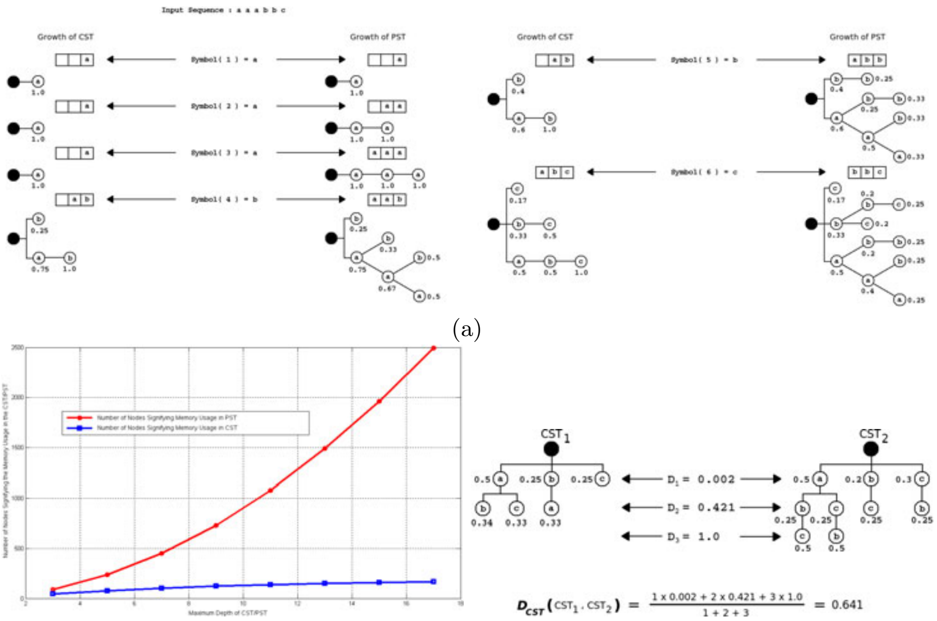
The process of sequence mining initializes with an empty (first in first out) buffer $\beta(0)$ (of length L, the maximum allowable sequence length) and the null tree $\mathcal{T}(0)$ (containing only the root node ρ). At the $(t-1)^{th}$ instant $(t > L)$, the buffer will contain the most recent L symbols from the sequence $\mathbf{S}$, i.e. $\beta(t-1) = \{\epsilon(t-L), \ldots, \epsilon(t-2), \epsilon(t-1)\}$. As the next symbol $\epsilon(t)$ is pushed to the buffer, the oldest symbol in the buffer is thrown out, i.e. $\beta(t) = \{\epsilon(t-L+1), \ldots \epsilon(t-1), \epsilon(t)\}$. Thus, the buffer (β) acts as a moving window (of size L) over the symbol stream $\mathbf{S}$. In case of a PST, the symbol $\epsilon(t)$ is directly pushed to the buffer β. However, in the proposed approach of transition sequence mining for learning a CST, $\epsilon(t)$ is pushed to β only in case of a transition, i.e. $\epsilon(t) \neq \epsilon(t-1)$.

At every instant, the variable length sequences observed within the buffer are learned in $\mathcal{T}$. For example, if a buffer β of size 4 contain symbols as $\beta = \{\ a,\ a,\ b,\ c\ \}$, then the 4 variable length sequences c, $b\ c$, $a\ b\ c$ and $a\ a\ b\ c$ are learned as branches in $\mathcal{T}$. Let the set of l-length paths (originating from ρ) of the tree $\mathcal{T}(t)$ at the t^{th} instant be $B^{(l)}(t) = \{\alpha_u^{(l)}(t), u = 1, \ldots b_l\}$, where b_l is the number of l-length branches in the tree. Now, if the l-length sequence in the buffer matches with the b^{th} path of $B^{(l)}(t)$, then the probabilities $\{\pi_u^{(l)}(t), u = 1, \ldots b_l\}$ of the nodes of $\mathcal{T}(t)$ at the l^{th} depth are updated as $\pi_u^{(l)}(t) = (1 - \eta_l(t))\pi_u^{(l)}(t-1) + \eta_l(t)\delta(u-b)$ where, $\eta_l(t)$ is the rate of learning l-length sequences at the t^{th} instant and $\delta(\bullet)$ is the Kronecker delta function. However, in the current implementation a fixed learning rate η is employed such that $\eta_l(t) = max\left(\frac{1}{t-l+1}, \eta\right) \forall l$.

The occurrence of a new symbol (with a transition in case of CST) results in the formation of newer variable length sequences in the buffer. Thus, new nodes signifying this symbol are added at different depths of the tree thereby growing newer branches. Each new node is initialized with an initial probability of $\eta_l(t)$, whereas the older node probabilities in the same depths are penalized by multiplying with a factor of $(1 - \eta_l(t))$. This ensures the self-normalizing nature of node probability updates such that they add up to 1.0 at each depth.

While learning these sequences, for the CST the probability of the 3 symbol sequence stays at 1.0 as it does not encounter any other transition sequence other than "$a\ b\ c$" till $t = 6$. On the other hand, the PST learns all the ordered sequences (of length 1/2/3 symbols) including the self-transitions. For example, at $t = 6$, the CST has only 6 nodes while the PST has 11 nodes. Thus, in practical applications of activity sequence modeling, where objects continue in the same state for long durations of time, the PST will learn a huge number of branches mostly due to self-transitions but the CST will remain compact in size for learning only branches with symbol-transitions. The memory efficiency of the CST is however reflected for higher values of L and longer sequences.

We illustrate this by performing (transition) sequence mining on 100 artificial symbol sequences of sizes between 900 – 1100. The symbols are drawn randomly from an alphabet of size 20. Any drawn symbol is repeated a random number of times (between 50 and 150) and the sequence is constructed accordingly. We learn both CST and PST from this sequence by varying the maximum depth of


Input Sequence : a a a b b c
Growth of CST
Growth of PST
Symbol(1) = a
Symbol(2) = a
Symbol(3) = a
Symbol(4) = b
Symbol(5) = b
Symbol(6) = c
Number of Nodes Signifying Memory Usage in PST
Number of Nodes Signifying Memory Usage in CST
Number of Nodes Signifying the Memory Usage in the CST/PST
Maximum Depth of CST/PST
2500
2000
1500
1000
500
0
2
4
6
8
10
12
14
16
18
CST1
CST2
D1 = 0.002
D2 = 0.421
D3 = 1.0
$D_{CST}(CST_1, CST_2) = \frac{1 \times 0.002 + 2 \times 0.421 + 3 \times 1.0}{1+2+3} = 0.641$

(a)

(b)

(c)

Fig. 2. (a) Learning the CST and the PST with a maximum depth of 3 from an input sequence of 6 symbols – "*a a a b b c*". The first three symbols being the same, the CST learns only a single node ("*a*" , 1.0) whereas the PST learns 3 nodes, i.e. ("*a*", 1.0), ("*a a*" , 1.0) and ("*a a a*" , 1.0). A transition is registered by the CST at $t = 4$ (new symbol: "*b*") and it learns two single symbol sequences ("*a*",0.75) and ("*b*",0.25) (*a* occurred thrice and *b* once in four symbols) and one 2-length sequence ("*a b*", 1.0). However, the PST learns two single symbol sequences ("*a*",0.75) and ("*b*",0.25); two sequences of size 2, i.e. ("*a a*", 0.67) and ("*a b*", 0.33); and, two sequences of length 3, i.e. "*a a a*" and "*a a b*" with probabilities 0.5 each. The learning process continues in the similar manner till $t = 6$. Note that, at $t = 6$, the CST has only 6 nodes while the PST has 11 nodes. (b) Memory usage in CST and PST for different tree depths. The maximum depth of the trees are varied from 3 to 17 in steps of 2. As the maximum depth of the tree increases, the number of nodes grow almost linearly in the CST as compared to the polynomial growth in case of the PST. (c) Illustrating the computation of distance between CSTs.

the tree (L) from 3 to 17 in steps of 2. The number of nodes added to the trees signify their memory usage during the process of (transition) sequence mining. Figure 2(b) shows the average number of nodes added to the CST/PST while the maximum depth of the tree is varied. As the maximum depth of the trees increase, the number of nodes grow almost linearly in the CST as compared to the polynomial growth in the case of PST. Notably, the CST presents a memory efficient alternative to the PST while mining sequences with large number of self-transitions as is frequently observed in the area of activity analysis.

Note that, the CST preserves the distribution of the symbols in the input, i.e. the probability distributions at the first depth of both CST and PST are the same and encodes the information relevant to transitions while using much lesser memory as compared to the PST. However, the PST maintains certain stochastic properties of the input sequences like variable order conditional probabilities and thus its branches can also be used as variable length symbol predictors. However, none of the earlier unsupervised activity analysis approaches have used these predictor properties as they were of no use in the context of activity structure discovery [2].

These stochastic properties are certainly lost in the CST as during the process of learning only transitions we are effectively editing the input sequence by removing redundant repetitions. However, we still maintain the relative frequencies of the states themselves and the self-transitionless sub-sequences. For example, in case of our input sequence "$a\ a\ a\ b\ b\ c$", we do not perform a buffer update operation at $t = 5$ when the symbol "b" repeats (Figure 2(a)), but we learn the fact that the transition sequence "$a\ b$" has occurred 2 times. This information is relevant in discriminating sequences with similar symbol transition structures but different symbol repetition durations. For example, the sequences "$a\ a\ a\ b\ b\ c$" and "$a\ a\ b\ b\ b\ c\ c$" have similar transition structure $a \rightarrow b \rightarrow c$ but different symbol repetitions. The procedure for computing (dis)similarities between symbol sequences modeled as CST are described next.

3 CST Clustering

We have proposed to learn CST from object state sequences where each tree as a whole characterizes each activity. To cluster these activities, we need to define a measure between two CSTs to reflect the dissimilarity between the activities that they have modeled. Motivated by the fact that the CST hosts probability distributions at each of its depth, we propose to use the Bhattacharya distance between two probability distributions $\mathcal{P}_1, \mathcal{P}_2$, given by $\mathcal{D}_b(\mathcal{P}_1, \mathcal{P}_2) = 1 - \sum_i \sqrt{\mathcal{P}_1(i)\mathcal{P}_2(i)}$. We define the dissimilarity $\mathcal{D}_{cst}(\mathcal{T}_1, \mathcal{T}_2)$ between the two CSTs $\mathcal{T}_1$ and $\mathcal{T}_2$ as $\mathcal{D}_{cst}(\mathcal{T}_1, \mathcal{T}_2) = \frac{\sum_{l=1}^{L} g(l)\mathcal{D}_b(\mathcal{P}_1^{(l)}, \mathcal{P}_2^{(l)})}{\sum_{l=1}^{L} g(l)}$ where $\mathcal{P}_k^{(l)}$ is the probability distribution hosted at the l^{th} depth of the k^{th} tree ($k = 1, 2$) and $g(l)$ is a monotonically increasing positive function of the depth l (in our case $g(l) = l$). The proposed dissimilarity measure (Figure 2(c)) exploits the distributions of all variable length subsequences of a temporal pattern and we choose to assign more weight to the longer sub-sequences as compared to the shorter ones to give more importance to the structure of the whole event rather than it's minute details (shorter sub-sequences). However, note that $\mathcal{D}_{cst}$ is only a semi-metric as it does not satisfy the triangular inequality. CSTs learned from time indexed object state sequences are further grouped using average linkage hierarchical clustering algorithm whose performance depends on the number of clusters. We next present the methodology adopted for computing the clustering sensitivity.

3.1 Clustering Performance Analysis

Consider the case of clustering N activities belonging to M different categories, such that the m^{th} category contain N_m activity instances ($N = \sum_{m=1}^{M} N_m$). Consider an unsupervised classification algorithm to form Q clusters over these activities. We can form a $M \times Q$ "*cluster-category distribution matrix*" ($CCDM$) such that, $CCDM[m][q]$ ($m = 1, \ldots M$, $q = 1, \ldots Q$) denotes the number of activities of the m^{th} category which belong to the q^{th} cluster. The category label $l(q)$ of the q^{th} cluster is assigned based on the maxima of the frequencies of the activity categories present in it. Thus, if instances of the i^{th} category are present with a maximum frequency in the q^{th} cluster, then $l(q) = i = argmax_j CCDM[j][q]$. This makes us to consider all the other activities in the q^{th} cluster which do not belong to $l(q) = i$ as false detections with respect to the i^{th} category.

A $M \times M$ confusion matrix (CM) of (unsupervised) classification can be constructed from $CCDM$ using the cluster-category labels. Let, $CM[r][k]$ ($r, k = 1, \ldots M$) denote the number of activities actually belonging to the r^{th} class, which have been classified to be belonging to the k^{th} category. It can be shown that the confusion matrix entries can be computed as $CM[r][k] = \sum_{q=1}^{Q} CCDM[r][l(q) - k]$. The sensitivity $S(Q)$ of unsupervised classification for Q number of clusters is defined as the fraction of the total number of activities which are classified correctly and is thus computed as $S(Q) = \frac{\sum_{r=1}^{M} CM[r][r]}{N}$. Sensitivity curves of activity categorization can be obtained by varying the desired number of clusters Q in the hierarchical agglomerative clustering.

4 Results

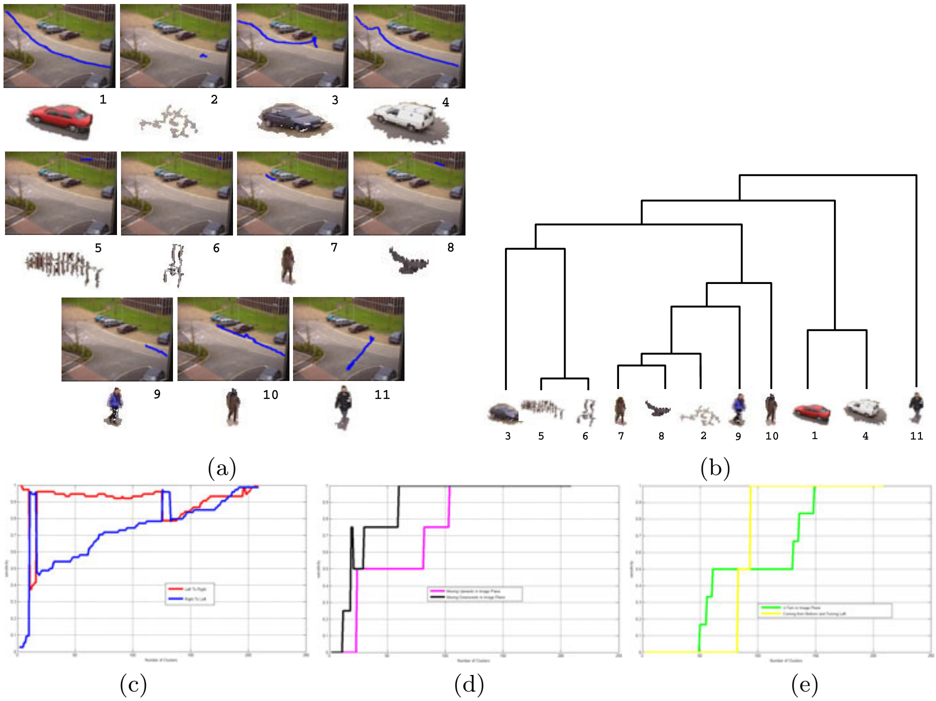
We present our results on two outdoor surveillance data sets – first, the PETS2000 data set and second, a traffic video data set (Figure 1). We illustrate our results mainly on the PETS2000 data set as it contains a smaller number of objects and show only performance analysis results on the traffic video.

Pixel-wise mixture of Gaussian based scene background modeling is used to detect the scene objects as foreground blobs. Object features (color distribution and motion model) learned from these blobs are used to track multiple objects across the images. We have used the multi-object tracking algorithm proposed in [10] to extract the object trajectories[1].

For the purpose of characterizing object actions in terms of the image data, we use a characterization of the image plane motion of the object (presented below). The motion directions in the image plane are quantized to form 8 motion states as in compass directions – M_1 denoting "eastwards", M_2 signifying "north-east" and so on going anti-clockwise to M_8 representing the "south-east" direction.

[1] Due to space constraints, we do not provide the details of foreground extraction and multiple object tracking algorithm for trajectory extraction.

Additionally, we use the state M_0 in case the object is at rest. Object trajectories in image space form an important aspect of single object actions. We represent the trajectory as a sequence of quantized motion direction states. CSTs are learned over the motion direction sequences of the objects which form the motion behavior model. The CSTs are further subjected to average linkage hierarchical agglomerative clustering algorithm for discovering action categories.



(a) (b)

(c) (d) (e)

Fig. 3. Trajectory clustering – (a) Trajectories of objects extracted from PETS2000 data set. Note that the noise objects #2 , #5, #6 and persons #7, #9 have very short scene presences and insignificant trajectories. (b) The dendogram formed by hierarchical agglomerative clustering of the CST trees learned from these trajectories. Note that the trajectories of persons #7, #9 and #10 are distinct from those of the cars #1 and #4. However, Car #3's parking trajectory distinguishes it from the other cars. (c)–(e) Trajectory clustering sensitivity vs. number of clusters (traffic video). For clarity, the graphs are shown for two image plane trajectory categories at a time – (c) LEFT TO RIGHT and RIGHT TO LEFT; (d) MOVING UPWARDS and MOVING DOWNWARDS; (e) U-TURN and COMING FROM BOTTOM AND TURNING LEFT.

Foreground blob detection followed by multiple object tracking is used to extract 11 objects from the PETS2000 data set whose trajectories are shown in figure 3(a). The dendogram formed by clustering the CST learned from quantized motion direction state sequences (obtained from object trajectories) is shown in figure 3(b). The red and white colored cars (objects #1 and #4) had almost similar trajectory which entered the scene from the right and drove to exit through the left boundary are clustered properly. Also, the objects #7, #8, #2, #9 and #10 had their trajectories directed from left to right albeit with varying frequencies and are seen to cluster in the same group. Also, it is worth noting that the trajectory (walking downwards in the scene and turning back) of object #11 (person in black dress) remained salient from the others in the process of clustering.

The processes of blob detection and tracking are used to extract 209 objects from the traffic video. The image plane trajectories of these objects characterize their actions. Manual inspection of these image plane trajectories show the existence of 6 different categories along with spurious ones (MISC) arising due to track losses – Γ(TRAJECTORY) = { LEFT TO RIGHT (75), RIGHT TO LEFT (74), MOVING UPWARDS (4), MOVING DOWNWARDS (4), U-TURN (3), COMING FROM BOTTOM AND TURNING LEFT (5), MISC (44) }. CSTs are learned from the temporally ordered sequences of the quantized motion directions obtained from the trajectories which are further subjected to hierarchical agglomerative clustering. The classification sensitivities of each category are computed by the performance analysis procedure outlined in sub-section 3.1. The classification sensitivities for varying number of clusters are shown in figures 3(c)–(e).

Note that the sensitivity computed by our evaluation criterion increases with the number of clusters in hierarchical agglomerative clustering (figure 3(c)–(e)). However, the trajectory categories of LEFT TO RIGHT and RIGHT TO LEFT appear as the leading or most frequent members of the clusters (Sub-section 3.1) with even lower number of clusters. Also, the trajectories of the categories MOVING UPWARDS and MOVING DOWNWARDS show high sensitivities at lower number of clusters even with lower frequency of appearance. We believe that the trajectories MOVING UPWARDS and MOVING DOWNWARDS stand out as they do not have motion states in common with LEFT TO RIGHT or RIGHT TO LEFT. On the other hand, the trajectories of U-TURN and COMING FROM BOTTOM AND TURNING LEFT show low sensitivities on account of high sub-structural similarities with the other four trajectory categories.

5 Conclusion

We have proposed to model activity descriptor sequences using "*compressed suffix trees*" (CST) which are shown to be advantageous over the existing formulations of using the "*probabilistic suffix trees*" (PST). This has direct advantages in memory efficient model construction thereby avoiding the learning of large number of self-transitions present in activity descriptor sequences. It is shown that the use of CST is advantageous in cases of higher order activity modeling as the CST shows an almost linear growth in memory usage while the PST grows polynomially in size (number of nodes) as the maximum allowable sequence length is increased. We have further proposed a distance metric in the space of CSTs to perform hierarchical agglomerative clustering of activity models. The efficiency of the proposed approach is experimentally verified on 2 data sets where activities are represented as quantized object motion direction state sequences.

Activities are generally described as temporally ordered sequences of the object states. However, another informative component of an activity is the number and nature of the participants in that event. Our future work aims at characterizing activities through the participant category (e.g. humans ride a bike), number of participants (group activity – people boarding a bus; single object action – running, walking etc. and two-object interactions – handshake, following, overtaking etc.) along with the state space features (pose/velocity/proximity etc.) of the participants. Discovery of such information beyond the analysis of only activity structures (i.e. models learned from time indexed object descriptor sequence) in a purely unsupervised framework will advance us a few steps further towards the key goals of a cognitive vision system.

References

1. Moeslund, T., Hilton, A., Kruger, V.: A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding 104, 90–126 (2006)
2. Galata, A., Johnson, N., Hogg, D.: Learning variable-length markov models of behavior. Computer Vision and Image Understanding 81, 398–413 (2001)
3. Gao, J., Hauptmann, A.G., Bharucha, A., Wactlar, H.D.: Dining activity analysis using a hidden markov model. In: IEEE International Conference on Pattern Recognition, pp. 915–918 (2004)
4. Bobick, A., Wilson, A.: A state-based technique for summarization and recognition of gesture. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 382–388 (1995)
5. Brand, M., Oliver, N., Pentland, A.: Coupled hidden markov models for complex action recognition. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1–6 (1997)
6. Johnson, N., Hogg, D.: Learning the distribution of object trajectories for event recognition. In: British Machine Vision Conference, pp. 583–592 (1995)
7. Buxton, H.: Learning and understanding dynamic scene activity: a review. Image and Vision Computing 21, 125–136 (2003)
8. Veeraraghavan, H., Papanikolopoulos, N., Schrater, P.: Learning dynamic event descriptions in image sequences. In: IEEE International Conference on Computer Vision and Pattern Recognition (2007)
9. Hamid, R., Maddi, S., Bobick, A., Essa, M.: Structure from statistics - unsupervised activity analysis using suffix trees. In: IEEE International Conference on Computer Vision (2007)
10. Guha, P., Mukerjee, A., Venkatesh, K.S.: Efficient occlusion handling for multiple agent tracking with surveillance event primitives. In: Second Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (2005)

A Continuous Learning in a Changing Environment

Aldo Franco Dragoni, Germano Vallesi, and Paola Baldassarri

DIIGA, Universitá Politecnica delle Marche,
via Brecce Bianche 1, 60131 Ancona, Italia
{a.f.dragoni,g.vallesi,p.baldassarri}@univpm.it

Abstract. We propose a Hybrid System for dynamic environments, where a "Multiple Neural Networks" system works with Bayes Rule to solve a face recognition problem. One or more neural nets may no longer be able to properly operate, due to partial changes in some of the characteristics of the individuals. For this purpose, we assume that each expert network has a reliability factor that can be dynamically re-evaluated on the ground of the global recognition operated by the overall group. Since the net's degree of reliability is defined as the probability that the net is giving the desired output, in case of conflicts between the outputs of the various nets the re-evaluation of their degrees of reliability can be simply performed on the basis of the Bayes Rule. The new vector of reliability will be used to establish who is the conflict winner, making the final choice (the name of subject). Moreover the network disagreed with the group and specialized to recognize the changed characteristic of the subject will be retrained and then forced to correctly recognize the subject. Then the system is subjected to continuous learning.

Keywords: Belief revision, face recognition, neural networks, unsupervised learning, hybrid system.

1 Introduction

A single neural network cannot effectively solve some complex problems as several researches in the field of Artificial Neural Networks show[1]. This led to the concept of "Multiple Neural Networks" systems for tackling complex tasks improving performances w.r.t. single network systems [2]. The idea is to decompose a large problem into a number of subproblems and then to combine the individual solutions to the subproblems into a solution to the original one [2]. This modular approach can lead to systems in which the integration of expert modules can result in solving problems which otherwise would not have been possible using a single neural network [3]. The responses of the individual modules are simple and have to be combined by some integrating mechanism in order to generate the complex overall system response [4]. The combination of individual responses is particularly critical when there are incompatibilities between them. Such situations may arise for example when the system operates

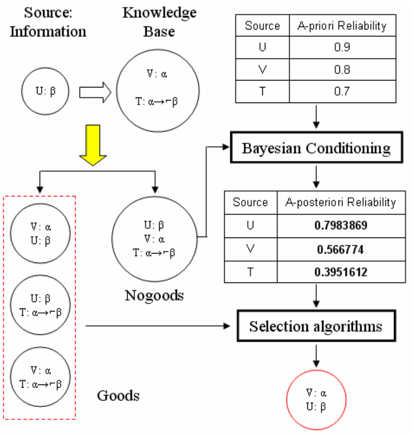
G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 79–88, 2011.

in dynamic environments, where it can happen that one or more modules of the system are no longer able to properly operate [5].

In this context, we propose a "Multiple Neural Networks" system to solve a face recognition problem. Each part of the system consisted of a single neural network is trained to recognize a specific region of the face and to each one is assigned an arbitrary a-priori reliability. Each network has an a-priori reliability factor that is the likelihood that the source is considered credible. This factor will be dynamically re-evaluated on the ground of the global recognition operated by the overall group. In other words, in case of conflicts between the outputs of the various nets the re-evaluation of their "degrees of reliability" can be simply performed on the basis of the Bayes Rule. The conflicts depend on the fact that there may be no global agreement about the recognized subject, may be for she/he changed some features of her/his face. The new vector of reliability obtained through the Bayes Rule will be used for making the final choice, by applying the "Inclusion based" algorithm [3] or another "Weighted" algorithm over all the maximally consistent subsets of the global output of the neural networks. The nets recognized as responsible for the conflicts will be automatically forced to learn about the changes in the individuals characteristics through a continuous learning process.

2 Theoretical Background

In this section we introduce some theoretical background taken from the Belief Revision (BR) field. Belief Revision occurs when a new piece of information inconsistent with the present belief set (or database) is added in order to produce a new consistent belief system [6].



Source	A-priori Reliability
U	0.9
V	0.8
T	0.7

Source	A-posteriori Reliability
U	**0.7983869**
V	**0.566774**
T	**0.3951612**

Fig. 1. "Belief Revision" mechanism

In figure 1, we see a Knowledge Base (KB) which contains two pieces of information: the α information , which come from V source, and the rule "If α, then not β" that comes from T source. Unfortunately, another piece of β information produced by the U source , is coming, causing a conflicts in the KB. To solve the conflicts we have to found all the "maximally consistent subsets", called Goods, inside the inconsistent KB, and we choose one of them as the most believable one. In our case (figure 1) there are three Goods: $\{\alpha, \beta\}$; $\{\beta, \alpha \rightarrow \neg\beta\}$; $\{\alpha, \alpha \rightarrow \neg\beta\}$. Maximally consistent subsets (Goods) and minimally inconsistent subsets (Nogoods) are dual notions. Given an inconsistent KB finding all the Goods and finding all the Nogoods are dual processes. Each source of information is associated with an a-priori "degree of reliability", which is intended as the a-priori probability that the source provides correct information. In case of conflicts the "degree of reliability" of the involved sources should decrease after "Bayesian Conditioning" which is obtained as follows. Let $S = \{s_1, ..., s_n\}$ be the set of the sources, each source s_i is associated with an a-priori reliability $R(s_i)$. Let ϕ be an element of 2^S. If the sources are independent, the probability that only the sources belonging to the subset $\phi \subseteq S$ are reliable is:

$$R(\phi) = \prod_{s_i \in \phi} R(s_i) * \prod_{s_i \notin \phi} (1 - R(s_i)) \tag{1}$$

This combined reliability can be calculated for any ϕ providing that:

$$\sum_{\phi \in 2^S} R(\phi) = 1 \tag{2}$$

Of course, if the sources belonging to a certain ϕ give inconsistent information, then $R(\phi)$ must be zero. Having already found all the Nogoods, what we have to do is:

- Summing up into $R_{Contradictory}$ the a-priori reliability of Nogoods;
- Putting at zero the reliabilities of all the contradictory sets, which are the Nogoods and their supersets;
- Dividing the reliability of all the other (no-contradictory) set of sources by $1 - R_{Contradictory}$ we obtain the new reliability (NR).

The last step assures that the equation 2 is still satisfied and it is well known as "Bayesian Conditioning". The revised reliability $NR(s_i)$ of a source s_i is the sum of the reliabilities of the elements of 2^S that contain s_i. If a source has been involved in some contradictions, then $NR(s_i) \leq R(s_i)$, otherwise $NR(s_i) = R(s_i)$.

For instance, the application of this Bayesian conditioning to the case of Figure 1 is showed in the following table 1 and table 2.

2.1 Selection Algorithms

These new or revised "degrees of reliability" will be used for choosing the most credible Good as the one suggested by "the most reliable sources". One of the

Table 1. Conflict table

ϕ	R(U)	R(V)	R(T)	$R(\phi)$	$NR(\phi)$
	0.1	0.2	0.3	0.006	0.0120967
T	0.1	0.2	0.7	0.014	0.0282258
V	0.1	0.8	0.3	0.024	0.048387
VT	0.1	0.8	0.7	0.056	0.1129032
U	0.9	0.2	0.3	0.054	0.1088709
UT	0.9	0.2	0.7	0.126	0.2540322
UV	0.9	0.8	0.3	0.216	0.4354838
UVT	0.9	0.8	0.7	0.504	
				$\sum_{\phi \in 2^S} R(\phi) = 1$	$\sum_{\phi \in 2^S} NR(\phi) = 1$

Table 2. Revised reliability

ϕ	$NR(\phi)$	$NR(U \in S)$	$NR(V \in S)$	$NR(T \in S)$
	0.0120967	0	0	0
T	0.0282258	0	0	0.0282258
V	0.048387	0	0.048387	0
VT	0.1129032	0	0.1129032	0.1129032
U	0.1088709	0.1088709	0	0
UT	0.2540322	0.2540322	0	0.2540322
UV	0.4354838	0.4354838	0.4354838	0
UVT	0	0	0	0
		NR(U)=0.7983869	NR(V)=0.566774	NR(T)=0.3951612

algorithms to perform this job is called "Inclusion based" (IB) [7]. This algorithm works as follows:

1. Select all the Goods which contains information provided by the most reliable source;
2. if the selection returns only one Good, STOP, that's the searched most credible Good;
3. else, if there are more than one Good then pop the most reliable source from the list and go to step 1;
4. if there are no more Goods in the selection, the ones that were selected at the previous iteration will be returned as the most credible ones with the same degree of credibility.

The other algorithm is "Inclusion based weighted" (IBW) a variation of Inclusion based: each Good is associated with a weight derived from the sum of Euclidean distances between the neurons of the networks (i.e. the inverse of the credibility of the recognition operated by each net). If IB selects more than one Good, then IBW selects as winner the Good with a lower weight.

The third and also the last algorithm is "Weighted algorithm" (WA) that combines the aposteriori reliability of each network with the order of the answers provided. Each answer has a weight $1/n$ where $n \in [1; N]$ represents its position among the N responses. Every Good is given a weight obtained by joining together the reliability of each network that supports it with the weight of the answer given by the network itself, as shown in the following equation 3:

$$W_{Good_j} = \sum_{i=1}^{M_j} (\frac{1}{n_i * Rel_i}) \tag{3}$$

where W_{Good_j} is the Weight of $Good_j$; Rel_i is the reliability of the network i; n_i is the position in the list of answers provided by the network i and finally M_j is the number of network that compose $Good_j$. If there are more than one Good with the same reliability then the winner is the Good with the highest weight.

3 Face Recognition System: An Example

In the present work, to solve the face recognition problem [8], we use a "Multiple Neural Networks" system consisted of a number of independent modules, such as neural networks, specialized to recognize individual template of the face. We use 4 neural networks specialized to perform a specific task: eyes (E), nose (N), mouth (M) and, finally, hair (H) recognition. Their outputs are the recognized subjects, and conflicts are simple disagreements regarding the subject recognized. As an example, lets suppose that during the testing phase, the system has to recognize the face of four persons: Andrea (A), Franco (F), Lucia (L) and Paolo (P), and that, after the testing phase, the outputs of the networks are as follows: E gives as output "A or F", N gives "A or P", M gives "L or P" and H gives "L or A"', so the 4 networks do not globally agree. Starting from an undifferentiated a-priori reliability factor of 0.9, and applying the Belief revision method, for each expert network we get the following new degree of reliability: NR(E) = 0.7684, NR(N) = 0.8375, NR(M) = 0.1459 and NR(H)=0.8375. The networks N and H have the same reliability, and by applying a selection algorithm it turns out that the most credible Good is {E,N,H}, which corresponds to Andrea. So Andrea is the response of the system.

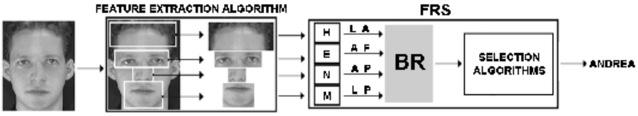


Fig. 2. Face Recognition System (FRS) representation

Figure 2 shows a schematic representation of the Face Recognition System (FRS), which is able to recognize the most probable individual even if there are serious conflicts among their outputs.

4 Face Recognition System in a Dynamical Environment

As seen in the previous section, one or more networks may fail to recognize the subject, there can be two reasons for the fault of the net: either the task of recognizing is objectively harder, or the subject could have recently changed something in the appearance of his face (perhaps because of the grown of a goatee or moustaches). The second case is very interesting because it shows how our FRS could be useful for implementing Multiple Neural Networks able to follow dynamic changes in the features of the subjects. In a such dynamic environment, where the input pattern partially changes, some neural networks could no longer be able to recognize the input. However, if the changes are minimal, we guess that most of the networks will still correctly recognize the face. So, we force the faulting network to re-train itself on the basis of the recognition made by the overall group. Considering the a-posteriori reliability and the Goods, our idea is to automatically re-train the networks that did not agree with the others. The network that do not support the most credible Good is forced to re-train themself in order to "correctly" (according to the opinion of the group) recognize the face. Each iteration of the cycle applies Bayesian conditioning to the a-priori "degrees of reliability" producing an a-posteriori vector of reliability. To take into account the history of the responses that came from each network, we maintain an "average vectors of reliability" produced at each recognition, always starting from the a-priori degrees of reliability. This average vector will be given as input to the two algorithms, IBW and WA, instead of the a-posteriori vector of reliability produced in the current recognition. In other words, the difference with respect to the BR mechanism is that we do not give an a-posteriori vector of reliability to the two algorithms (IBW and WA), but the average vector of reliability calculated since the FRS started to work with that set of subjects to recognize. Now the subject has moustaches and goatee, while, when the system is trained, the subject did not have them. So O_M network (specialized to recognize the mouth) is no longer able to correctly indicate the tested subject. Since all the others still recognize Andrea, O_M will be retrained with the mouth of Andrea as new input pattern.

The re-learning procedure occurs when the changing is longer than the previously fixed temporal window (windowlength equals to 10) associated to each neural network. So we avoid the re-learning for a subject with a very variable feature. We define imm_i the portion of the image containing the feature

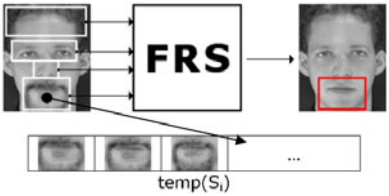


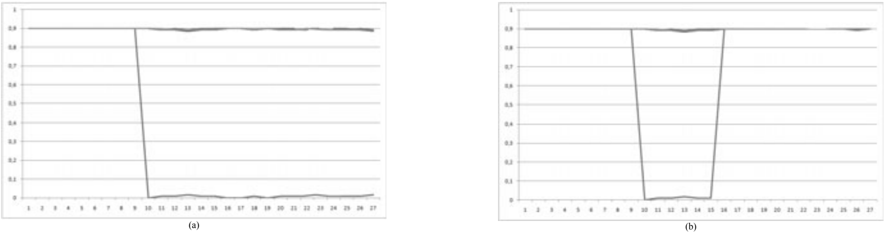
Fig. 3. Functioning of the temporal window

analyzed by the network r_i; S the subject identified by the synthesis function of the FRS; s_{ik} is the subject i in the k-th position of the list ordered on the base of the distance of the LVQ output. So the re-learning procedure consists of the following steps:

1. For each network r_i the system compares S and s_{ik} used to find the Good. If $S \neq s_{ik} \forall k$ then in the temporary directory $temp(S_i)$ (that is the temporal window) of the subject S related to the network i is saved the imm_i portion, as showed in Figure 3. On the contrary if $S = s_{ik}$ for one k the temporary directory temp(S_i) is empitied;
2. If in temp(S_i) there are windowlength samples, the $temp(S_i)$ images are transferred in riadd(S_i) removing its old images, then the retraining of the r_i network begins using the riadd(S_i) images for S and the most recent images for all other subjects.

We have to highlight that the windowlength chosen strongly depends on the variability of subjects, and so on the database used for the testing. It is important also to note that there will always be a limit to the size of the windowlength beyond which for any dataset the system will be able to filter all the changes, to a value beyond this limit the system behaves as a system without re-learning. If not recognized by the networks, the introduction of re-learning in the facial recognition system allows, that the networks maintain higher reliability values than in the case without re-learning, as shown in Figures 4a and 4b. This because, the network now can recognizes a feature that could not recognize with the original knowledge acquired during the first training of all networks. If a network is no longer able to recognize one feature can not contribute to the final choice. Moreover if other networks do not recognize the subject, but indicate the same wrong subject, the whole system fails. In this case there would be a wrong Good that could be the most likely for the system but associated to the incorrect subject.

Figure 4 shows the a-posteriori reliability trend related to five expert neural networks concerning a particular subject. Observing the two graphs, we can see that until the networks are agree, the reliability maintains high values. While, if one of the networks (eg mouth) comes into conflict with the others giving in output another subject (since perhaps he changed his appearance) then the



(a)

(b)

Fig. 4. Performance of the a-priori reliability (a) without and (b) with re-learning

reliability goes down. In Figure 4a, we can see how this conflict will bring the net loser to have a low reliability. Conversely, in Figure 4b, we can see that if the network does not recognize the subject for a consecutive number of times corresponding to the windowlength samples the re-learning begins, after which the a-posteriori reliability again increases.

5 Experimental Results

This section shows only partial results: those obtained without the feedback, discussed in the previous section. In this work we compared two groups of neural networks: the first consisting of four networks and the second of five networks (the additional network is obtained by separating the eyes in two distinctive networks). All the networks are LVQ 2.1, a variation of Kohonens LVQ [9], each one specialized to respond to individual template of the face. The training set is composed of 20 subjects (taken from FERET database [10]), for each one 4 pictures were taken for a total of 80. Networks were trained, during the learning phase, with three different epochs: 3000, 4000 and 5000. To find Goods and Nogoods, from the networks responses we use two methods:

1. Static method: the cardinality of the response provided by each net is fixed a priori. We choose values from 1 to 5, 1 meaning the most probable individual, while 5 meaning the most five probable subjects.
2. Dynamic method: the cardinality of the response provided by each net changes dynamically according to the minimum number of "desired" Goods to be searched among. In other words, we set the number of desired Goods and reduce the cardinality of the response (from 5 down to 1) till we eventually reach that number (of course, if all the nets agree in their first name there will be only one Goods).

In the next step we applied the Bayesian conditioning depending from Goods obtained with these two techniques. In this way we obtain the new reliability for each network. These new "degrees of reliability" will be used for choosing the most credible Good (then the name of subject). We use two selection algorithms to perform this task: Inclusion based weighted (IBW), Weighted algorithm (WA). To test our work, we have taken 488 different images from 20 subjects and with these images we have created the Test set. As shown in figure 5 using the system without re-learning the results show how the union of the Dynamic method with the selection algorithm WA and five neural networks gives the best solution to reach a 79.39% correct recognition rate of the subjects. Moreover using only one LVQ network for the entire face, we obtain the worst result. In other words, if we consider a single neural network to recognize the face, rather one for the nose, one for the mouth and so on, we have the lowest rate of recognition equals to 66%. This is because a single change in one part of the face makes the whole image not recognizable to a single network, unlike a hybrid system.

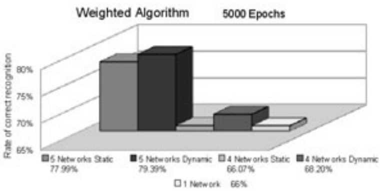


Fig. 5. Average rate of correct recognition with either Test Set and the results obtained using only one network for the entire face

In Figure 6, we can see the comparison between the average rate of correct recognition in the following cases:

– Hybrid system with re-learning (Static method, WA selection algorithm), 89.25%
– Hybrid system without re-learning (Static method, WA selection algorithm), 79.39%
– Only one neural network for the entire face, 66%.

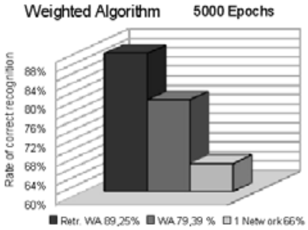


Fig. 6. Average recognition with re-learning

So the re-learning is a procedure very useful not only to increase the reliability factor of the misleading network, but also to improve the recognition itself.

6 Conclusion

Our hybrid method integrates multiple neural networks with a symbolic approach to Belief Revision to deal with pattern recognition problems that require the cooperation of multiple neural networks specialized to recognize subjects that dynamically change some their characteristics for that some nets occasionally fail. We tested this hybrid method referring to a face recognition problem, training each network to recognize a specific region of the face: eyes, nose, mouth, and hair. Every output unit is associated with one of the persons to be recognized. Each net gives the same number of outputs. We consider a constrained environment in which the image of the face is always frontal, lighting conditions, scaling and rotation of the face being the same, as in all biometric recognition systems for access to restricted areas. We accommodated the test so that changes of the faces are partial, for example the mouth and hair do not change simultaneously, but one at a time. Under this assumption of limited changes, our hybrid system ensures great robustness to the recognition. Will never happen that an authorized person tries to access a restricted area controlled by a biometric recognition systems with his face distorted. In case of permanent injury of face all the networks will be retrained to the new face by the system operator. The system assigns a reliability factor to each neural network, which is recalculated on the basis of conflicts that occur in the choice of the subject. The new "degrees of reliability" are obtained through the conflicts table and Bayesian Conditioning. These new "degrees of reliability" can be used to select the most likely subject. When the subject partially changes its appearance, the network responsible for the recognition of the modified region comes into conflict with other networks and its degree of reliability will suffer a sharp decrease. So, the overall system is engaged in a never ending loop of testing and re-training that makes it able to cope with dynamic partial changes in the features of the subjects. To maintain high values of the reliability for all the networks is very important since the choice of the right subject strongly depends on the credibility of all the experts.

References

1. Azam, F.: Biologically inspired modular neural networks. PhD Dissertation, Virginia Tech. (2000)
2. Shields, M.W., Casey, M.C.: A theoretical framework for multiple neural network systems. Neurocomputing 71(7-9), 1462–1476 (2008)
3. Sharkey, A.J.: Modularity combining and artificial neural nets. Connection Science 9(1), 3–10 (1997)
4. Li, Y., Zhang, D.: Modular Neural Networks and Their Applications in Biometrics. Trends in Neural Computation 35, 337–365 (2007)
5. Guo, H., Shi, W., Deng, Y.: Evaluating sensor reliability in classification problems based on evidence theory. IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics 36(5), 970–981 (2006)
6. Gardenfors, P.: Belief Revision. In: Cambridge Tracts in Theoretical Computer Science, vol. 29 (2003)
7. Benferhat, S., Cayrol, C., Dubois, D., Lang, J., Prade, H.: Inconsistency management and prioritized syntax-based entailment. In: Proceedings of the 13th International Joint Conference on Artificial Intelligence, pp. 640–645 (1993)
8. Tolba, A.S., El-Baz, A.H., El-Harby, A.A.: Face Recognition a Literature Review. International Journal of Signal Processing 2, 88–103 (2006)
9. Kohonen, T.: Learning vector quantization. In: Self Organizing Maps. Springer Series in Information Sciences, Berlin, vol. 30 (2001)
10. Philips, P.J., Wechsler, H., Huang, J., Rauss, P.: The FERET Database and Evaluation Procedure for Face-Recognition Algorithms. Images and Vision Computing Journal 16(5), 295–306 (1998)

Human-Computer Interaction through Time-of-Flight and RGB Cameras

Piercarlo Dondi, Luca Lombardi, and Marco Porta

Department of Computer Engineering and Systems Science, University of Pavia,
Via Ferrata 1, 27100 Pavia, Italy
{piercarlo.dondi,luca.lombardi,marco.porta}@unipv.it

Abstract. The number of systems exploiting Time-of-Flight (ToF) cameras for gesture recognition has greatly increased in the last years, confirming a very positive trend of this technology within the field of Human-Computer Interaction. In this work we present a new kind of application for the interaction with a virtual keyboard which is based on the use of an ordinary RGB webcam and a ToF camera. Our approach can be subdivided into two steps: firstly a segmentation of the entire body of the user is achieved exploiting only the ToF data; then the extraction of hands and head is obtained applying color information on the retrieved clusters. The final tracking step, based on the Kalman filter, is able to recognize the chosen hand also in presence of a second hand or the head. Tests, carried out with users of different ages, showed interesting results and a quick learning curve.

Keywords: Time-of-Flight camera, human-computer interaction, hand recognition.

1 Introduction

Time-of-Flight (ToF) cameras are able to measure depth in real-time using a single compact sensor, unlike previous multi-camera systems, such as stereo cams. In the last years, research has shown a large interest in such devices in many fields related to computer vision and computer graphics, like 3D modeling, scene reconstruction, user interaction or segmentation and tracking of moving people [1]. In some cases, the ToF contribution is combined with color informations supplied by a traditional RGB camera to achieve more complex or precise results: for instance, in [2] depth and color data are used to create a 3D ambient for mixed reality; in [3] depth information is exploited to select the best input area for a color-based segmentation algorithm (SIOX); while in [4] a fusion of colors and depth data is employed in a new segmentation and tracking method for compensating the respective weaknesses of the two different kinds of sensors.

In this paper we focus on the combined use of a RGB and a ToF camera for Human-Computer Interaction (HCI), presenting a new application which allows the user to control virtual on-screen keyboards (in particular a QWERTY keyboard and a numeric pad). The ToF stream is used for the initial search of the

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 89–98, 2011.

entire body of the user, using an approach described in a previous work [5]; this solution achieves a real-time foreground segmentation which is robust to sunlight noise. Color information is then applied to the retrieved clusters to extract head and hands. The subsequent tracking step, implemented using a Kalman filter, is able to follow the hand of interest also when the other hand is detected or in presence of overlaps with the head. Experimental tests, carried out with users of both sexes aged between 20 and 50, showed interesting results: in particular a quick learning curve and a fast decrease of errors.

The paper is organized as follows: section 2 provides an overview of the main characteristics of ToF cameras; section 3 presents the state of the art related to HCI applications based on these devices; section 4 describes the proposed method; section 5 shows the experimental results; section 6, at last, draws some conclusions.

2 Time-of-Flight Cameras

Cameras based on the Time-of-Flight principle work in the near infrared band exploiting laser light to assess distances of elements in the scene, thus providing depth information. Practically, two major approaches exist to implement these devices, namely pulsed and modulated light. In the first case the target is hit by a coherent wavefront and the depth is measured analyzing the variations of the reflected wave; in the second case an incoherent modulated light is used and the time of flight is determined by means of phase delay detection.

Compared to stereo cameras or laser scanners, ToF cameras are characterized by some benefits: they do not employ moving mechanical components, work reliably in real-time, are not affected by shadows and can calculate 3D distances within any scenario. A drawback of these systems is however their sensitivity to sun light, which introduces considerable noise (on the contrary, in general, artificial illumination does not interact with the sensor). Although a ToF camera has a nominal working range of about 10 m, noise caused by scattering, multipaths and environment light may decrease this value, and the actual range is therefore between two and five meters [6].

To implement the system described in this paper we exploited the SR3000 ToF camera by MESA Imaging [7], a modulated-light device which we used at 20MHz and whose active sources work in the near infrared, at about 850nm. Its frame rate is about 18-20 fps. The camera produces two maps per frame – each with a resolution of 176x144 pixels – one containing distance information and the other reporting the intensity of light reflected by objects in the scene. Since the sensor is not affected by visible light, values of intensity depend on energy in the near infrared range only. For this reason, nearer objects look clearer, since they reflect more light, while more distant elements appear darker.

3 Previous Works

The number of systems exploiting the ToF approach for gesture recognition has greatly increased in the last years, confirming a very positive trend of this technology within the field of Human-Computer Interaction. In almost all implementations, the first step is the removal of background objects by means of a threshold applied to the assessed distance between hand/arm and the camera. Then, the next stages depend on the specific task.

The system described in [8], for example, carries out gesture recognition by fusing 2D and 3D images. The segmentation technique employed is based on the combination of two unsupervised clustering approaches,K-Means and Expectation Maximization, which both try to locate the centers of natural clusters in the combined data. Another gesture recognition system based on hand movement detection is presented in [9] After rejecting elements falling outside a predefined depth range, the Principle Component Analysis (PCA) technique is exploited to get a first basic estimation of the position and orientation of the hand. Afterward, a more complex (3D skeleton-based) hand model is matched to the previously acquired data. In [10], 12 static gestures are classified according to X and Y projections of the image and depth information. The projections of the hand are used as features for the classification, while the arm area is removed. Depth features are taken into account to distinguish gestures which have the same projections but different alignments. The technique described in [11] combines a pre-trained skin color model (a Gaussian mixture approach) with a histogram-based adaptive model which is updated dynamically with color information extracted from the face. The system has been tested with sample images containing six different hand postures and can identify gestures and movements of both hands.

Slideshow control is a typical application of gesture recognition. As an example in [12], the "thumbs-up" gestures towards left or right are used to switch to the previous or next slide, while pointing to the screen is interpreted as a "virtual laser pointer". The pointing direction is calculated in 3D, at first through a segmentation of the person in front of the camera with respect to the background, and then by detecting the 3D coordinates of head and hand. An analogous use for the interaction with a beamer projector is described in [7].

A further context for effective gesture recognition is represented by medical applications, where it is sometimes necessary a touchless interaction: an example is proposed in [13], where a system based on ToF camera allows the exploration and navigation through 3-D medical image data.

Usage scenarios for the ToF approach are however not limited to those quoted above, but can be extended to several settings. A very comprehensive survey of developments of ToF technology and related applications can be found in [1].

4 Feature Detection and Interaction

4.1 Foreground Segmentation

Our segmentation algorithm [5] is designed so as not to need any preprocessing operations or a priori knowledge of the environment or of the shapes of objects. This section summarizes its main steps comprehensive of noise compensations; section 4.2 describes the proposed color based extension for hand recognition.

Two steps make up our approach: a first thresholding of the distance map based on the corresponding values in the intensity image, followed by a region growing stage that starts from seeds planted on peaks of the intensity map. Considering the characteristic of the ToF camera described in section 2, we use intensity map as a guide to restrict the area of investigation in the range map and to find good candidates to become seeds.

For every frame we dinamically estimate a proper intensity threshold (λ_{seed}) applying the Otsu's method. This parameter is used to define the set of seeds S (1).

$$S = \{P_x : I_x > \lambda_{seed}, \|P_x - P_s\| > \gamma, \gamma > 1\} \tag{1}$$

P_x is a point of the distance map, I_x is its corresponding intensity value and P_s is the last seed found. The presence of a control of the distance between seeds guarantees their better distribution and reduces significantly their number in order to decrease the time needed for the following growing step.

The similarity measure S between a cluster pixel x and a neighboring pixel y is defined in (2):

$$S(x, y) = |\mu_x - D_y| \tag{2}$$

D_y is the distance value of pixel y and μ_x is a local parameter related to the mean distance value around x (6). The lower S is, the more similar the pixels are. When a seed is planted, μ_x is initialized to D_x. Considering a 4-connected neighborhood, a pixel x belonging to a cluster C absorbs a neighbor y according to the following conditions:

$$\{x \in C, S(x, y) < \theta, I_y \in L, \theta > 512\} \rightarrow \{y \in C\} \tag{3}$$

where L is the set of points of intensity generated using the equations (4) and (5) designed to threshold the data compensating the effects of noise caused by sunlight:

$$A = \{I_y : (I_y > \lambda) \vee [(I_y < \lambda) \wedge (I_{8n} > \delta * \lambda)], \delta \in [0, 1], \lambda < \lambda_{seed}\} \tag{4}$$

$$L = A \cup M \tag{5}$$

where λ is an intensity threshold proportional to λ_{seed}, I_{8n} is the intensity of all the neighbors of the pixel y considering the 8-connection, and M is the set A after the application of a series of morphological operations experimentally established (in order, two dilations, five erosions and a final dilation).

When a neighbor y of seed x is absorbed, we compute the average distance value μ_y in an incremental manner as follows:

$$\mu_y = \frac{\mu_x * \alpha + D_y}{\alpha + 1} \tag{6}$$

Parameter α is a learning factor of the local mean of D. If pixel y has exactly α neighbors in the cluster, and if the mean of D in this neighbor is exactly μ_x, then μ_y becomes the mean of D when y is added to the cluster. Every region grows excluding the just analyzed pixels from successive steps. The process is iterated for all seeds in order of descending intensity. Regions too small, with dimension lower than a fixed value, are discarded.

An experimental evaluation of the performances on different kinds of computers (both desktops and notebooks), made in a previous work [5], showed that the proposed approach ensures a good compromise between computational time and precision of the results: the system can reach the 44 fps with a high level computer and keeps the 18 fps of the ToF camera also with a low level one.

4.2 Hand Recognition

Hand detection is a complex task due to the high variability of hand shapes. An approach based only on color may be simpler but its performance is generally not good: objects in the background with color similar to skin produce inevitably false positives. This issue can be solved by limiting the region of interest: the described procedure retrieves the clusters in the foreground excluding automatically all the objects in the background. Moreover, considering the proposed interaction, in which the user must be relatively close to the camera to see what s/he is writing on the screen, we can further reduce the possible hand candidates, excluding a priori all the retrieved clusters placed too far from the camera (generally over 2 m).

So, after these preliminary phases we obtain a cluster of a half-body user (Fig. 1(a)) from which hand detection can start using the color information supplied by a standard webcam (in our experiments a Logitech HD Pro Webcam C910 with a resolution of 640x480 pixels). The calibration is achieved using a method similar to that described in [14].

Firstly we convert the image from the RGB to the HSV color model; then we eliminate all the points of the cluster that are outside the set W:

$$W = \{y : 0^\circ < H_y < 10^\circ, 350^\circ < H_y < 360^\circ, S_y > TH_S, V_y > TH_V\} \qquad (7)$$

where H_y, S_y and V_y are, respectively, the hue, saturation and value of the pixel y. The first two constraints define the color area with a hue in the skin range; the threshold on the saturation eliminates all the white points; finally, the search for points with high values of lightness excludes clothes with skin-like colors. For our purpose we do not need a precise segmentation of the hand but only an approximation, since the pointer on the keyboard is not controlled by the shape of the hand but by the position of its centroid (section 4.3). This simplification gives two advantages: we can use very strict conditions on color thresholding for finding the hand (we can afford to lose some details in order to remove certainly wrong areas) and the user can position his or her hand in the way s/he finds more comfortable. Small inaccuracies, like holes, are in any case fixed applying a morphological dilation on the retrieved sub-clusters.

For a better performance we apply this sub-segmentation procedure not after the entire foreground segmentation (section 4.1) but at the end of the thresholding phase (equations (4) and (5)): this way we can execute a single region growing procedure on a reduced set of points.

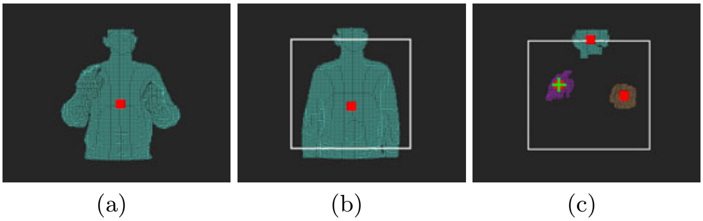


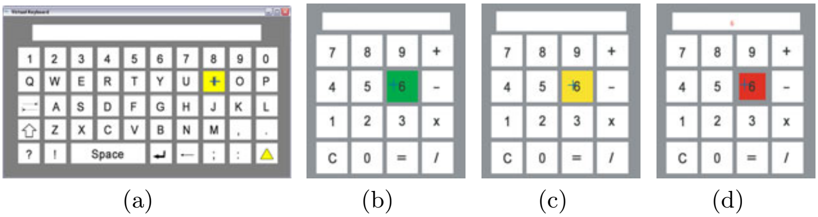
Fig. 1. Results of the segmentation steps, visualized as cloud of points: (a) initial segmentation of the entire body; (b) search for the active area (white rectangle); (c) hands and head extraction – the cross points the active hand

The last issue to solve is the choice of the active cluster and the exclusion of the others. We designed a training phase in which the user must stay in front of the camera and raise the hand that s/he has chosen to use. The system easily distinguishes between the hand (the cluster closer to the camera) and the head (the cluster in the upper position). For following the hand in the next iterations we use a tracking approach based on Kalman filter. The association between measured clusters and Kalman trackers is evaluated by minimum square Euclidean distance between the centroid of each cluster and the position predicted by each Kalman. This method, described in [5], is able to track multiple subjects, also in presence of short-term occlusions, and can thus be successfully applied in this situation. The use of the tracker allows the chosen hand to be recognized also in presence of other moving clusters like a second hand (Fig. 1(c)).

4.3 Interaction with Keyboard/Numeric Pad

We created for our application two kinds of keyboard with keys of different sizes: a reduced QWERTY keyboard (Fig. 2(a)) and a numeric pad (Fig. 2(b)). The interaction with the two keyboards occurs the same way in both cases: moving the hand moves the cursor pointer. Once the user has chosen the key s/he wants to press, s/he needs only to move the hand towards the camera, like if virtually pushing the key (the ToF camera can measure variations in distance of the tracked hand with no computational overhead). We set an appropriate distance threshold (experimentally determined) beyond which the key is considered pressed (section 5).

To estimate the position of the hand, we use its centroid, because it is the most stable point in the cluster: errors in depth evalutation caused by motion artifacts or by sunlight mainly affect points on the edges of the objects. The system performs a mapping of the position of the hand in the camera frames to the position of the cursor on the keyboard, but not all the area framed is considered valid. In fact, if the corners of the visual field of the camera corresponded to the corners of the keyboard, it would be very uncomfortable for the user to select and to press the keys placed in those positions. After specific tests, it was found that the more comfortable ″interaction zone″ for the user is a vertical



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

Fig. 2. The two kinds of keyboard used: (a) the reduced QWERTY (800x450 pixels on a 1680x1050 screen); (b)(c)(d) the numeric pad (550x600) with the three possible key states: (b) green - the pointer is on the key, but the key is not yet selected; (c) yellow - the key is selected, ready to be pressed; (d) red - the key is pressed

rectangle limited horizontally by the shoulders and vertically by the head and stomach (Fig. 1(b)). The selection of the active area for each user is automatically determinated during the training step.

Some visual feedbacks help the user to understand the state of the system. A key becomes green when the pointer is on it (Fig. 2(b)). A key becomes yellow (Fig. 2(c)), and is considered selected and ready to be pressed, when the pointer remains on it for a short period of time (around 1 sec). This status is useful to avoid accidental presses. Finally a key becomes red and gets smaller when it is pressed and the corresponding character appears in the box on top of the keyboard (Fig. 2(d)).

5 Experimental Results

A set of experiments were carried out to test the system and to obtain learning curve of the input method. 16 users participated in the tests (8 males and 8 females) aged between 20 and 50 (mean 28). All of them were placed in front of the cameras at a distance so that they were framed at half-body. Before starting the tests, each user was briefly trained in writing with the two keyboards (for about 2 minutes).

In the first test we asked the user to write the word *"ciao"* (*"hello"* in Italian) with the QWERTY keyboard; the purpose was to find the best press threshold in terms of time spent, mistakes made and personal preferences. The threshold value is the minimum distance (in cm) from the ToF camera below which a key is considered pressed. The results show that most of the users chose as the best threshold the nearest one (ThN = 50 cm) or the medium one (ThM = 60 cm). This is reasonable because the farthest threshold (ThF = 75 cm) is too much sensitive and it is easier to push a key unintentionally. Table 1 shows the writing times obtained for the three thresholds, as well as the errors made by each tester. The mean values were 16.2, 15.9 and 12.7 seconds, respectively, for ThN, ThM and ThF. A within-subjects ANOVA did not find any evident connection between threshold and times ($F = 3.03$, $p = .058$). A clear relation

Table 1. Test 1: writing a word. Time required and errors with different thresholds (ThN = 50cm, ThM = 60cm, ThF = 75cm). Highlighted cells indicate the threshold preferred by each user.

	ThN		*ThM*		*ThF*	
	Time (sec)	*Errors*	*Time (sec)*	*Errors*	*Time (sec)*	*Errors*
Tester 1	18	0	20	0	13	2
Tester 2	13	0	10	0	10	2
Tester 3	10	1	10	0	9	3
Tester 4	13	0	20	0	9	2
Tester 5	19	1	14	3	10	3
Tester 6	19	0	17	0	15	0
Tester 7	15	1	15	1	13	3
Tester 8	13	0	10	0	10	2
Tester 9	13	1	12	1	18	4
Tester 10	19	0	23	0	9	2
Tester 11	22	1	19	0	13	0
Tester 12	21	1	13	0	14	2
Tester 13	14	0	21	0	13	2
Tester 14	12	0	9	1	14	2
Tester 15	8	2	9	2	10	1
Tester 16	16	0	17	0	15	0

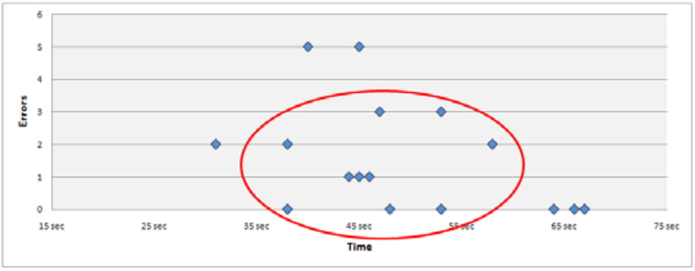


Fig. 3. Test2: writing a sentence with the threshold chosen in first test. Each point corresponds to a user

emerged instead between threshold and errors ($F = 9.79$, $p < .001$), with many more mistakes made with ThF.

In the second test we asked the users to write the sentence *"Ciao, come stai?"* (*"Hello, how are you?"* in Italian) using the thresholds chosen in the first test. The results obtained were interesting because they provided an indication about the user learning curve after a short period of system use. The plot in figure 3 shows a good balance between time and errors: for most users (see the area surrounded by the red circle) the task took between 35 and 60 seconds to complete, with a number of errors quite similar to that of the first part of the experiment (single word writing).

Table 2. Test 3: making a calculation

	ThN		*ThM*		*ThF*	
	Time (sec)	*Errors*	*Time (sec)*	*Errors*	*Time (sec)*	*Errors*
Tester 1	10	0	9	0	12	2
Tester 2	18	0	15	1	12	1
Tester 3	10	0	11	0	15	3
Tester 4	16	1	11	0	12	2
Tester 5	12	0	12	0	7	1
Tester 6	22	0	19	0	18	0
Tester 7	21	1	15	0	20	0
Tester 8	33	0	18	0	13	0
Tester 9	22	0	13	1	13	2
Tester 10	13	0	10	0	12	0
Tester 11	20	0	16	0	10	0
Tester 12	19	0	11	0	10	0
Tester 13	14	1	12	0	13	2
Tester 14	14	1	16	0	11	1
Tester 15	15	1	13	1	20	2
Tester 16	16	0	17	0	15	0

Finally, we carried out an experiment with the numeric pad (the user had to make the calculation $58 + 32$). Similarly to the first experiment, the purpose was to find whether there were changes in threshold preferences with larger keys (their size was twice that of the keyboard's keys). Like in the first experiment the thresholds were tested in randomized order. The results show a predictable reduction of errors (close to zero) with ThN and ThM, while with ThF their number was as high as in the first test. It is interesting to note that, whereas in the first experiment the difference in times with the three thresholds was minimal, now ThF is significantly greater. This anomaly may be explained considering that with small keys the great part of the time is spent selecting the correct key, while with big keys the selection is faster and the time required to press the key becomes more relevant. These considerations explain the nearly unanimous choice of the ThM threshold. The mean values for times were 17.2, 13.6 and 13.3 seconds, respectively, for ThN, ThM and ThF. A within-subjects ANOVA found clear relations both between threshold and times ($F = 4.04, p < .05$), with longer times with ThN, and between threshold and errors ($F = 6.3$, $p < .005$), with many more mistakes with ThF.

6 Conclusions

In this paper we have presented a new kind of gestural interaction with virtual keyboards that exploits the potentials of the combination of an RGB and a ToF camera. The system is totally independent of the background and of the shape of the hand, and the tracking stage enables the continuous identification of the active hand also in presence of similar clusters, such as a second hand or part of an arm. The experimental evaluation performed with 16 users showed interesting results, in particular a rapid learning curve and a sensible reduction of mistakes after few minutes. Future improvements include a more precise color sub-segmentation for excluding possible false positives in the hand detection step and the concurrent use of two hands for writing faster.

References

1. Kolb, A., Barth, E., Koch, R., Larsen, R.: Time-of-Flight Cameras in Computer Graphics. Computer Graphics Forum 29, 141–159 (2010)
2. Bartczak, B., Schiller, I., Beder, C., Koch, R.: Integration of a Time-of-Flight camera into a mixed reality system for handling dynamic scenes, moving viewpoints and occlusions in real-time. In: Fourth International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT 2008) (2008)
3. Santrac, N., Friedland, G., Rojas, R.: High resolution segmentation with a time-of-flight 3d-camera using the example of a lecture scene. Technical report (2006), http://www.inf.fu-berlin.de/inst/agki/eng/index.html
4. Bleiweiss, A., Werman, M.: Real-time foreground segmentation via range and color imaging. In: Kolb, A., Koch, R. (eds.) Dyn3D 2009. LNCS, vol. 5742, pp. 58–69. Springer, Heidelberg (2009)
5. Dondi, P., Lombardi, L.: Fast Real-Time Segmentation and Tracking of Multiple Subjects by Time-of-Flight Camera. In: 6th International Conference on Computer Vision Theory and Applications (VISAPP 2011), pp. 582–587 (2011)
6. Oprisescu, S., Falie, D., Ciuc, M., Buzuloiu, V.: Measurements with ToF Cameras and Their Necessary Corrections. In: International Symposium on Signals, Circuits and Systems, ISSCS 2007 (2007)
7. Oggier, T., Büttgen, B., Lustenberger, F., Becker, G., Rüegg, B., Hodac, A.: Swissranger SR3000 and First Experiences based on Miniaturized 3D-TOF Cameras. In: Proceedings, 1st Range Imaging Research Day, September 8-9, pp. 97–108. ETH Zurich Switzerland (2005)
8. Ghobadi, S., Loepprich, O., Hartmann, K., Loffeld, O.: Hand Segmentation Using 2D/3D Images. In: Image and Vision Computing, New Zealand, pp. 64–69 (2007)
9. Breuer, P., Eckes, C., Müller, S.: Hand Gesture Recognition with a Novel IR Time-of-Flight Range Camera–A Pilot Study. In: Gagalowicz, A., Philips, W. (eds.) MIRAGE 2007. LNCS, vol. 4418, pp. 247–260. Springer, Heidelberg (2007)
10. Kollorz, E., Penne, J., Hornegger, J., Barke, A.: Gesture recognition with a Time-Of-Flight camera. Int. J. Intell. Syst. Technol. Appl. 5(3/4), 334–343 (2008)
11. Van den Bergh, M., Van Gool, L.: Combining RGB and ToF cameras for real-time 3D hand gesture interaction. In: IEEE Workshop on Applications of Computer Vision (WACV 2011), pp. 66–72 (2011)
12. Haker, M., Böhme, M., Martinetz, T., Barth, E.: Deictic Gestures with a Time-of-Flight Camera. In: Kopp, S., Wachsmuth, I. (eds.) GW 2009. LNCS, vol. 5934, pp. 110–121. Springer, Heidelberg (2010)
13. Soutschek, S., Penne, J., Hornegger, J., Kornhuber, J.: 3-D Gesture-Based Scene Navigation in Medical Imaging Applications Using Time-Of-Flight Cameras. In: IEEE Computer Vision and Pattern Recognition Workshops, pp. 1–6 (2008)
14. Reulke, R.: Combination of distance data with high resolution images. In: Image Engeeniring and Vision Metrology, IEVM 2006 (2006)

Handling Complex Events in Surveillance Tasks

Daniele Bartocci and Marco Ferretti

Department of Computer Engineering and Systems Science
University of Pavia, Italy
daniele.bartocci@consorzio-cini.it, marco.ferretti@unipv.it

Abstract. In this paper we fully develop a fall detection application that focuses on *complex event* detection. We use a decoupled approach, whereby the definition of events and of their complexity is fully detached from low and intermediate image processing level. We focus on context independence and flexibility to allow the reuse of existing approaches on recognition task. We build on existing proposals based on domain knowledge representation through ontologies. We encode knowledge at the rule level, thus providing a more flexible way to handle complexity of events involving more actors and rich time relationships. We obtained positive results from an experimental dataset of 22 recordings, including simple and complex fall events.

Keywords: fall detection, complex event, rule engine.

1 Introduction

The automatic detection of events in video stream is one of the most promising applications of computer vision. Its application field includes surveillance, monitoring, human-computer interaction and content-based video annotation.

Event detection systems usually follow these major steps [1]: extraction of low-level features, recognition of actions or primitive events, high-level semantic interpretation of the events occurring.

Over the past decade many approaches have been proposed to perform the first two tasks (see [2] for a survey) which are often context dependent. Research on the latter stage of event detection is instead leaning towards description-based approaches to ensure independence from the application domain.

Recent works on this topic suggest the use of ontologies and rule languages to detach domain knowledge from the application code. The VERL language [3] has been proposed to define events, but it lacks a full implementation.

Snidaro et al. [4] used the standardized Ontology Web Language (OWL) and Semantic Web Rule Language (SWRL) to represent and maintain domain knowledge in surveillance applications. They further explored this approach [5] by replacing SWRL with a dedicated rule language, Jess. Their experimental architecture allows to infer events given an input of primitive events, and to define the system behavior when an event triggers.

We designed our framework for event detection on top of a similar architecture, which decouples image processing from reasoning. We also adopted their domain

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 99–108, 2011.

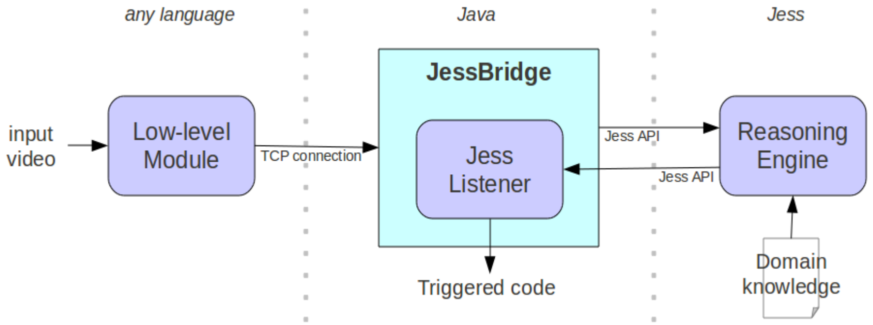


Fig. 1. Architecture overview

knowledge organization but propose a different approach in encoding to better address the definition and detection of *complex events*. Our goal is framework evaluation in terms of flexibility and complex event handling.

In order to achieve this goal we developed a complete application instance in the fall detection application field, tailoring a promising approach by Rougier et al. [6] to suit our framework. Potential falls are detected through Motion History Image (MHI) and shape analysis; the reasoning module distinguishes confirmed falls from false alarms and identifies *complex collective* falls.

The rest of the paper is organized as follows. The framework architecture is discussed in the next section. Section 3 covers the encoding of domain knowledge. The test application is presented, along with experimental results, in sections 4 and 5. Finally, conclusions are drawn in section 6.

2 Framework Architecture

We designed the framework architecture to efficiently implement an event detection system that separates action recognition from domain knowledge and reasoning. This allows for easy reuse of previous approaches on recognition task, while using a dedicated rule engine to carry out the reasoning task.

We adopted the Jess rule engine, a complete environment with its own scripting language to define and evaluate rules. Jess can act as a standalone application but it also exposes a powerful API to embed the rule engine into any Java application.

These features make it suitable for a simple yet effective client/server architecture, based on TCP connections. The client (low-level module) implements all signal processing routines needed to recognize actions or primitive events from the input video stream, sending the results to the server.

The server is the reasoning module; it acts as an interface to the rule engine, feeding it with the messages received from the client and behaving accordingly to the events detected.

Figure 1 shows the overall framework architecture. The core component, called *JessBridge*, is written in Java language and fills the server role. Upon initialization it invokes a Jess engine, loading the domain knowledge from a configuration file.

When a connection is established with the client a buffer is used to store and forward incoming messages to Jess, which executes them. We assume that each row is a valid Jess command (stating an entity or action detected by the low-level module) or a system command (ex: connection break, engine re-initialization).

A separate Java class, which extends the *JessListener* interface (from Jess APIs), defines the system behavior. This class is the handler associated to the rule engine, and its methods will be invoked when an event is inferred or specific rules fire. Thanks to this handler we can execute any kind of triggered code: database updates, automated phone calls, activation of an alarm.

In terms of flexibility, the proposed architecture is platform independent and the whole reasoning task can be defined simply modifying the domain knowledge file and the listener class.

The architecture can be expanded by connecting many clients to the same rule engine using a separate thread for each TCP connection. This also allows the system to interact with other security devices such as smartcard readers and biometric sensors.

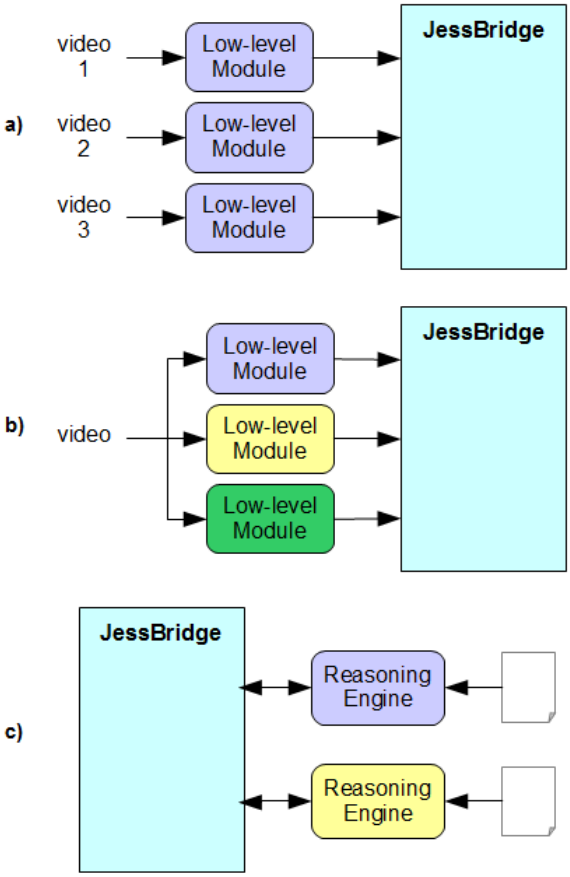


Fig. 2. Expanded architecture examples; *a)* multi-camera environment; *b)* different action classifiers; *c)* multiple rule engines

As shown in figure 2 each client may consist of the same low-level module connected to a different video stream (a common multi-camera environment). This way the reasoning module would be capable of inferring complex events based on simple events occurring at different locations. Otherwise the architecture may comprehend two or more dedicated low-level modules on the same stream, such as different action classifiers.

It is remarkable that the same server can handle more than one Jess instances, each with its very own domain knowledge. Such an architecture may be used for automatic rule learning and ontology evaluation.

3 Embedding Domain Knowledge

Domain knowledge is a key part of the event detection system. It ranges from the different types of entities that can be recognized to the relations and interactions that could take place between them. In our approach it also reflects in how the low-level module encodes its output. As discussed in [5], the knowledge needed by the reasoning engine can be split into the following categories, which differ in purpose and usage:

Static knowledge: taxonomic definitions of event detection's concepts (entities, actions and events). It may include some static instance of entities describing the passive environment such as: door, ATM, etc.
Dynamic knowledge: set of entities, actions and events detected or inferred by the system, each represented by a single instance of a concept in the static knowledge.
Rule knowledge: set of rules that define both simple and complex events for the application context.

Besides dynamic knowledge, contained in the rule engine working memory during the application execution, all definitions must be encoded (offline) in a proper way and loaded upon initialization. Unlike previous works we decided to encode taxonomical knowledge and rule knowledge in Jess language. Our different approach enables a simple solution to define complex events. Since common ontology languages (like OWL) can be easily translated in Jess language [5] if needed, this choice doesn't harm our flexibility goal.

In Jess terminology an instance is a *fact*, which is added dynamically to the working memory through an *assert* command. Each concept in the static knowledge is defined by a fact template with a list of *slot*s to express concept's properties; static instances can be added with the *deffacts* construct. The following example shows our basic template for the event concept:

```
(deftemplate event
    (slot type)
    (slot ID (default-dynamic (gensym*)))
    (multislot subjects)
    (slot t)     ;time
)
```

In this template, the `type` property is an instance of an event category, which can be defined as an enumeration or can be drawn from an ontology. The slot `ID` is system assigned when a fact is asserted in memory. This template can be used for *both* simple and complex events. The `multislot` property allows to associate to the event either a single or a list of values for event's `subjects`; this Jess feature is the key to handle *complex events* with a non-fixed number of subjects such as collective events. The `time` property is handled accordingly to the event category: in simple events (one subject only), it is the time at which the event is instantiated, in complex ones it can be suitably defined. In the following sections an example will be provided.

The *defrule* construct is used to specify the rule knowledge. Each rule has two parts: a pattern which is used to match facts in the working memory and the action to execute when the rule is triggered. Rules are usually composed of logical expressions that specify temporal and spatial constraints between entities and actions (or events). Instances in working memory are matched to verify these constraints. The activation of a simple rule can assert, modify or retract (delete from working memory) a fact concerning an entity or event. Examples of the *defrule* construct will be given and duly commented in section 4.2.

An appropriate flushing strategy has to be included in the rule set in order to manage instances' life cycles.

4 Test Application

To test our framework we developed a simple event detection system based on the proposed architecture. The steps required to deploy the complete application are limited to: low-level module implementation, specific domain knowledge design, definition of the listener class specifying the system behavior (in our test simply log messages). The application detects falls using an approach tailored on the work by Rougier et al. [6], as described in the next subsection. False alarms are filtered out by the reasoning module. Additionally the rule engine can identify and track collective falls.

4.1 Low-Level Module

Action recognition is achieved through the analysis of a mix of image features: a motion coefficient and the orientation and proportion of human shape. This module is implemented in C language using the OpenCV library.

First of all we need to identify moving people in the sequence. For this purpose the image is segmented with the background subtraction method described in [7], then each extracted blob is tracked between subsequent images using a predictor based on a Kalman filter, and processed individually.

We quantify motion using a coefficient based on the Motion History Image (MHI) H_τ of the blob, computed over 500ms:

$$C_{motion}(t) = \frac{\sum_{\forall(x,y)\in blob} H_\tau(x,y,t)}{\#\, pixel \in blob} \, . \tag{1}$$

Since motion is large when a fall occurs we filter out minor movements by thresholding C_{motion}. If a large motion is detected ($C_{motion} > 65\%$) we further analyze the image to discriminate a fall from other activities such as running. The blob is approximated by an ellipse using moments [8] calculating its orientation θ and semi-axis ratio ρ. We consider that a large motion is a fall if the standard deviation of one of these two features, computed for a duration of 1s, is over a fixed threshold (15 degrees on orientation and 0.9 on aspect ratio).

When a potential fall is detected a command is sent to the reasoning module to assert an *action:fall* fact. Furthermore the low-level module periodically notifies Jess whether the blob is moving (*action:move*), checking the ellipse's center coordinates.

4.2 Reasoning Module

In [6] false alarms and negligible falls are filtered out checking if the person is motionless on the ground after a possible fall. We detached this task from the image processing routines and implemented it in our reasoning module. In order to do so we refined the static domain knowledge to fit the application context.

The sole entity involved is the *agent*, a person who can perform the following actions: *move* or *fall*. The defined events include the simple event *confirmed-fall* and *recovery*, plus a complex *collective-fall* event. A *confirmed-fall* is an event that includes a fall, whereby the person involved remains still at ground, while a *recovery* event is a fall followed by a movement of the fallen subject, that actually recovers and moves away.

We provide a rule (depicted in figure 3) to infer an *event:recovery* when a movement is detected within 5 seconds after a fall – in Jess terms an *action:fall* and an *action:move* fact, with the same subject and appropriate time values, are found in the working memory. Likewise the complementary event (*confirmed-fall*) is inferred when no movement is detected.

We also used static instances to define inactivity zones, areas in which a fall is legitimate such as a bed or a couch. A dedicated rule with a higher priority than the previous two filters out those falls.

Lastly a complex event, *collective-fall*, is defined as two or more confirmed falls occurring in a short time span. Two rules are required to detect a collective event properly. The first asserts the event when the minimum constraints are satisfied – in our example two confirmed falls with different subjects and close time values (less than 6 seconds between falls), as expressed by the first two constraints of the rule:

```
(defrule collective_fall_create
    (event (type confirmed-fall)            ;1st constraint
           (subjects ?id1)
           (t ?t1))
    (event (type confirmed-fall)            ;2nd constraint
           (subjects ?id2&~?id1)
           (t ?t2&:(< (abs (- ?t1 ?t2)) 6)))

    (not (event  (type collective-fall)     ;3rd constraint
                 (subjects $? ?id1|?id2 $?)))
    =>
    (assert (event (type collective-fall)   ;rule effects
                   (subjects ?id1 ?id2)
                   (t (time))))
)
```

The two *confirmed-fall* facts hold a single subject value – namely `id1` and `id2` – since they are simple events, while the complex *collective-fall* event asserted as an effect of rule activation will hold the complete list of subjects involved. The third constraint avoids the assertion of a new instance of the complex event if another subject suffers a fall.

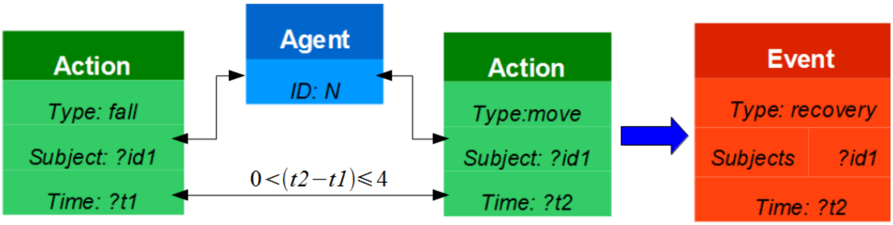


Fig. 3. Representation of the *event:recovery* inference rule

The second rule modifies the *collective-fall* event instance by expanding its subjects list whenever a new *confirmed-fall* event happens within a chosen time interval:

```
(defrule collective_fall_add
      (event (type confirmed-fall)            ;1st constraint
             (subjects ?idNew)
             (t ?tNew))
      ?e <-  (event (type collective-fall)    ;2nd constraint
             (subjects $?list&:(not (member$ ?idNew $?list)))
             (t ?t&:(< (abs (- ?t ?tNew)) 6)))
      =>
      (modify ?e (subjects $?list ?idNew))    ;rule effects
  )
```

The second constraint of the rule specifies that the new fall subject is matched against the *multislot* list to see if it is already in the list; if not, it is added if the temporal constraint is met. In the previous rule we defined the `time` property associated to the complex event as the current system timestamp. This means that the temporal constraint in the second rule will match any *confirmed-fall* asserted during 6 seconds following the *collective-fall* detection.

5 Experimental Results

To evaluate the test application, thereby the framework viability, we wanted to compare our approach on this subject with existing ones, such as [6]. We proceeded to record a dataset representing normal activities and simulated falls, including collective falls (missing in [6]). The application is designed to work with a single uncalibrated camera in a low-cost environment, so our video sequences were acquired using a simple USB webcam. Despite the low quality images (high compression artifacts and noise) we obtained positive results, as will be shown in the sequel.

The dataset (see Table 1) is composed of 22 sequences representing normal activities and simulated falls – harmful and recovered. The number of actors ranges from one (in 12 sequences) to four, exemplifying both simple and complex scenes. Figure 4 shows two snapshots taken from the recorded dataset. Obtained results are shown in table 2.

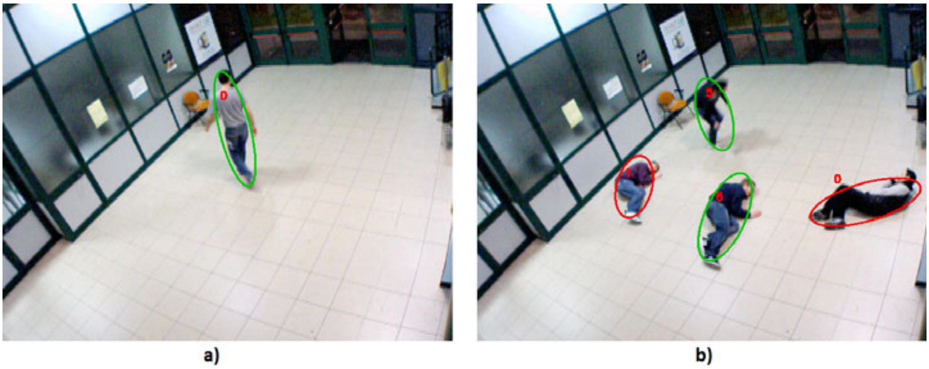


Fig. 4. Excerpts from the dataset; a) simple, single actor scene; b) complex, collective fall. Red ellipses represent already confirmed falls. The third person lying down will soon be tagged as "fallen" too.

Table 1. Dataset sequences

Sequence ID	Name	Agents	# frames	Confirmed falls	Recovery	Other actions
1	simple_fall_1	1	158	1		
2	recovery	1	158		1	
3	simple_fall_2	1	130	1		
4	simple_fall_3	1	113	1		
5	chair_fall_1	1	283	1		1
6	chair_fall_2	1	180	1		1
7	chair_liedown	1	203			2
8	crouching	1	184	1		1
9	liedown_recovery	1	199		1	1
10	pick_up_1	1	176	1		1
11	pick_up_2	1	155		1	2
12	spin_fall	1	223	1	1	
13	wandering	2	178	1		1
14	run_liedown	2	113		1	1
15	indifferent	2	136	1		1
16	run_away	2	81	1		
17	chair_fall_3	3	197	1		
18	collective_1	3	146	2		
19	collective_2	4	171	4		
20	collective_3	4	168	3	1	
21	collective_4	4	162	3	1	
22	hostile	4	171			2

Table 2. Fall detection results

	Detected	**Not detected**
Harmful falls	True positive: **15**	False negative: 2
Lures	False positive: 3	True negative: **19**

We get an overall good rate of event detection and false detection, with a sensitivity of 88.24% (ratio of true positive vs (true positive + false negative)) and a specificity of 86,36% (ratio of true negative vs (false positive + true negative)). Even if our dataset includes many sequences of higher complexity, our results are similar to those obtained by Rougier et al. [6]. We deduce that our decoupled approach, based on a rule engine, is suitable for event detection. Moreover our framework can reuse and expand previous approaches, without worsening the results.

Moreover, the collective fall event was included in 4 video sequences and detected in 3 of them, whenever the *confirmed-fall* events were correctly inferred. Due to complex blob intersections and subsequent tracking errors, the collective event could not be detected in one of those sequences, since only one out of three harmful falls were correctly recognized as *confirmed-fall*.

In summary, results show that we can correctly detect and handle complex events thanks to the rule engine but the higher the constraints complexity, the higher the chance of false negatives because of misclassifications and errors in the low-level module.

6 Conclusions

In this work we have shown an implementation of a complete framework for an event detection system which can handle complex events. We focused on flexibility and context independence. The approach is based on a client/server architecture which decouples action recognition from domain knowledge and reasoning. We deployed domain knowledge with a Jess-based rule engine, which performs the reasoning task.

We also developed a test application in the fall detection application field obtaining preliminary positive results and confirming the correctness of the approach. The experiments carried out emphasize the dependence of the reasoning task on a reliable recognition module. A more robust system could be obtained by strengthening the inference engine, accompanied by a probabilistic output from the recognition module. A possible approach is to extend Jess with fuzzy rule evaluation, thus modeling uncertainty in inferences using probabilistic ontologies.

References

1. Turaga, P., Chellappa, R., Subrahmaniam, V.S., Udrea, O.: Machine Recognition of Human Activities: A survey. IEEE Transactions on Circuits and Systems for Video Technology 18, 1473–1488 (2008)
2. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28(6), 976–990 (2010)

3. Francois, A.R., Nevatia, R., Hobbs, J., Bolles, R.C., Smith, J.: VERL: an ontology framework for representing and annotating video events. IEEE MultiMedia Magazine 12(4), 76–86 (2005)
4. Snidaro, L., Belluz, M., Foresti, G.L.: Representing and recognizing complex events in surveillance applications. In: Proc. of the 2007 IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 493–498 (2007)
5. Snidaro, L., Belluz, M., Foresti, G.L.: Modelling and Managing Domain Context for Automatic Surveillance Systems. In: Sixth IEEE International Conference on Advanced Video and Signal ased Surveillance (AVSS), pp. 238–243 (2009)
6. Rougier, C., Meunier, J., St-Arnaud, A., Rousseau, J.: Fall Detection from Human Shape and Motion History Using Video Surveillance. In: Proc. of the 21st International Conference on Advanced Information Networking and Applications Workshops, pp. 875–880 (2007)
7. Kim, K., Chalidabhongse, T., Harwood, D., Davis, L.: Real-time foreground-background segmentation using codebook model. Real-Time Imaging 11(3), 172–185 (2005)
8. Chaudhuri, B.B., Samanta, G.P.: Elliptic fit of objects in two and three dimensions by moment of inertia optimization. Pattern Recognition Letters 12(1), 1–7 (1991)

Face Analysis Using Curve Edge Maps

Francis Deboeverie[1], Peter Veelaert[2], and Wilfried Philips[1]

[1] Ghent University - Image Processing and Interpretation/IBBT,
St-Pietersnieuwstraat 41, B9000 Ghent, Belgium
Francis.Deboeverie@telin.ugent.be,
Wilfried.Philips@telin.ugent.be

[2] University College Ghent - Engineering Sciences,
Schoonmeersstraat 52, B9000 Ghent, Belgium
Peter.Veelaert@hogent.be

Abstract. This paper proposes an automatic and real-time system for face analysis, usable in visual communication applications. In this approach, faces are represented with Curve Edge Maps, which are collections of polynomial segments with a convex region. The segments are extracted from edge pixels using an adaptive incremental linear-time fitting algorithm, which is based on constructive polynomial fitting. The face analysis system considers face tracking, face recognition and facial feature detection, using Curve Edge Maps driven by histograms of intensities and histograms of relative positions. When applied to different face databases and video sequences, the average face recognition rate is 95.51%, the average facial feature detection rate is 91.92% and the accuracy in location of the facial features is 2.18% in terms of the size of the face, which is comparable with or better than the results in literature. However, our method has the advantages of simplicity, real-time performance and extensibility to the different aspects of face analysis, such as recognition of facial expressions and talking.

Keywords: geometric features, shape modeling, shape matching, face recognition, face tracking, facial feature detection.

1 Introduction

Automatic, reliable and fast face analysis is becoming an area of growing interest in computer vision research. This paper presents an automatic and real-time system for face analysis, which includes face recognition, face tracking and facial feature detection, and may be used in visual communication applications, such as video conferencing, virtual reality, human machine interaction, surveillance, etc.

The task of facial analysis has generally been addressed by algorithms that use shape and texture modelling. The Active Shape Model (ASM), proposed by Cootes et. al. [1], is one of the early approaches that attempts to fit the data with a model that can deform in ways consistent with a training set. The Active Appearance Model (AAM) [2] is a popular extension of the ASM. AAM is an integrated statistical model which combines a model of shape variation with a model of the appearance variations in a shape-normalized frame. The recently proposed Boosted Appearance Model (BAM), proposed by Liu et. al. [3,4], uses a shape representation similar to AAM, whereas

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 109–118, 2011.

the appearance is given by a set of discriminative features, trained to form a boosted classifier, able to distinguish between correct and incorrect face alignment. Liang et. al. [5] proposed a component-based discriminative approach for face alignment without requiring initialization. A set of learned direction classifiers guide the search of the configurations of facial components among multiple detected modes of facial components. Many of the methods mentioned are complex, dependent from initialization and not always capable to handle the different aspects of face analysis.

We propose face analysis with a compact feature, the Curve Edge Map (CEM), which describes polynomial segments with a convex region in an edge map. The face CEM is a simple and natural description, which still preserves sufficient information about the facial position and expression. In fact, many curve segments correspond to physically meaningful features such as eyebrows, cheekbones or lips. The CEM approach not only has the advantages of geometric feature-based approaches, such as low memory requirements, but also has the advantage of high recognition performance of template matching.

In this paper, we employ non-local matching of curve segments in different CEMs to find corresponding curve segments. For matching, the curve segments do not have to be in each others neighbourhood. The technique differs from the matching technique in [6], where a local matching technique is described using distance and intensity characteristics. The latter technique requires that CEMs are aligned. The matching technique we propose here, is more general and does not require that the CEMs are aligned. It considers three characteristics of the curve segments.

- The first characteristic is the orientation of the main axis of the curve segment to make a first binary classification.
- The second characteristic is the intensity difference between the inner and outer side of the curve segment. Each curve segment defines a convex region, so that it makes sense to distinguish curve segments from facial features with intensity histograms between the inner and outer side.
- The third characteristic is the relative position of the curve segment in the face CEM, found by the characteristic positions of the the curve segments from the facial features in the face CEMs. This topology is modeled by a histogram of relative positions in log-polar space, which is based on the work of shape contexts from Belongie et. al. [7].

The three characteristics classify an individual curve segment in the face CEM, which is useful for facial feature detection.

We give a brief overview of three tasks handled in our face analysis system, which will be explained in the following sections:

- *Face tracking:* The face CEMs in consecutive frames are matched using the non-local matching technique as mentioned above. The motion vectors from the corresponding curve segment pairs are constructed as described in [9], in which vehicles are succesfully tracked with parabola segments. Note that faces are first detected using the cascade-based face detection algorithm of Viola and Jones [8].
- *Face recognition:* The input face CEM is matched with the face CEMs from a database using the non-local matching technique. From the matching curve segment

pairs, a global matching cost is computed. A face is recognized when the global matching cost reaches a minimum.

- *Facial feature detection:* The curve segments from facial features, such as the eyebrows, the eyes, the nose and the lips, are detected in frontal face images, by matching the face CEMs with curve segment models. These curve segment models are built for each facial feature by a training proces and consist of the three characteristics as mentioned above.

Our method performs well when evaluated over different databases and video sequences, considering variations in scale, lighting, facial expressions and pose. The considered public available databases are the Georgia Tech Face Database [10], the Database of Faces [11], the Bern University Face Database [12], the AR Face Database [13], the Yale University Face Database [14] and the BioID Database [15] with ground truth marker points.

We achieve an average face recognition rate of 95.51%, which is better than the results for the technique described in [6]. Furthermore, we gain the advantage of facial feature detection, which automatically involves facial action recognition, such as recognition of facial expressions, talking and head movement. When applied to different test databases, the average facial feature detection rate is 91.92%. In this paper, we compare our results with the work of Cristinacce et. al. [16], in which they use a Constrained Local Model (CLM) to locate a set of feature points, initialized by facial feature locations found by applying the Viola and Jones face detector [8].

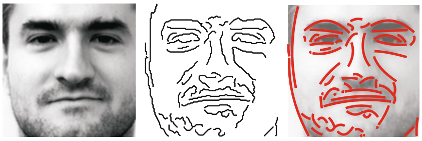
The accuracy in location of our facial features is computed for the BioID Database by comparing the ground truth marker points with the facial features detected. When using the mean euclidean distance as the basic error measurement, the mean location error is 6.53 pixels with standard deviation 1.15 pixels, which is 2.18% in terms of the size of the face. In this paper, we compare our results to the work of Ding et. al. [17], in which they achieve facial feature detection by learning the textural information and the context of the facial feature to be detected.

The face analysis results are comparable with or better than existing methods, when taking into account the limitations of image scale and lighting. However, the advantages of our method are simplicity, real-time properties and many face analysis tasks are handled by the same approach.

This paper is organized as follows. First, we briefly indicate how to compute a CEM for a gray level image in Section 2. In the following Section 3, we present the basic parts for matching the curve segments. Section 4 describes in detail the different applications of the face CEM. In Section 5 we show the results when testing our face analysis system on different databases and video sequences. Finally, we conclude our paper in Section 6.

2 Curve Segmentation

The basic step of our method is the interpretation and description of an image edge map by geometric primitives. Our feature representation, the Curve Edge Map (CEM), integrates structural information and spatial information by grouping pixels of an edge



(a) (b) (c)

Fig. 1. Curve segmentation: Images (a), (b) and (c) show the input image, the edge map and the face CEM, respectively.

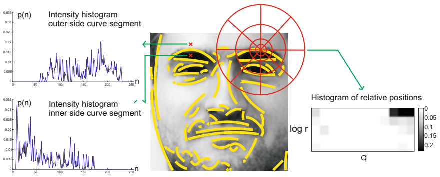


Fig. 2. Histograms of intensities and relative positions: On the left are shown the intensity histograms from the inner and outer side of the curve segment, which describes the left eyebrow. On the right is plotted the log-polar histogram of relative positions from the curve segment, which describes the right eyebrow.

map into curve segments. In [6,18], it is shown that the approximation of edge maps by second order polynomials or parabola segments is a useful and elegant representation for both non-rigid and rigid structures.

The edge map is computed with the Canny edge detector [19], a frequently used edge detector when considering edge information in faces [20], resulting in thin edges of one pixel thickness. Digitized curves are obtained from the edge map by a simple boundary scan algorithm. To fit curve segments, we use an adaptive incremental linear-time fitting algorithm for curve segmentation which is based on constructive polynomial fitting [21,18]. The output of the fitting algorithm is a list of curve segments that approximates the edge map with an L_∞ user-specified threshold. In practice, we represent the curve segments as parabola segments, represented by three parameters, which have their main axes in x or y direction, depending on which direction yields the smallest fitting cost. The restriction of allowing only two orientations is based on the many vertical and horizontal orientations of the facial features in a face. Figures 1 (a), (b) and (c) show the input image, the edge map and the CEM for a face from the Bern University Face Database [12], respectively.

3 Curve Edge Map Matching

3.1 Curve Coefficients Classification

The orientations of the curve axes are used to make a first binary classification, by which we mean that the curve segments will only match when their main axes have the same orientation. For example for the lips and the eyebrows the alignment of the curve axes are vertical, while for the nose the alignment of the curve axis is horizontal. In practice, we consider the axes of symmetry of the parabola segments, which are in the x or the y direction.

3.2 Curve Intensity Histograms

One of the discriminative properties in our system is the difference in intensity between the inner and outer side of the curve segment. For each side we construct a normalized intensity histogram.

When estimating a histogram from one side of the curve segment, the region of interest for intensities is in the area between the original curve segment and a duplicate which is translated parallel to the main axis of the curve segment.The distance d over which the duplicate is translated results in a robust face recognition rate R, when d is between 0.02 and 0.09 of the square root of the face surface A, where A is the rectangle produced by face detector. This optimization is done for the AR Face Database [13].

Figure 2 shows on the left two intensity histograms from the inner and outer side from the curve segment of the left eyebrow for a face from the Bern University Face Database [12]. The histogram representing the upper side has its intensities on the bright side, while the histogram representing the lower side has its intensities on the dark side.

To match the intensity histograms from one side of two different curve segments, we use the Bhattacharyya distance metric or B-distance measure, which measures the similarity of two probability distributions and is a value between 0 and 1. This measure is chosen for its good classification properties [22]. The B-distance between two intensity histograms f and g is

$$B(f,g) = 1 - \sum_{l=0}^{L} \sqrt{f(l)g(l)}, \tag{1}$$

where L is the number of bins in the histograms, $L = 255$ for gray images. The matching of intensity histograms is done for the inner and outer side of the curve segments, resulting in B_{in} and B_{out}.

3.3 Histograms of Relative Positions in CEM

A curve segment in the CEM is characterized by its relative position and becomes important when classifying individual facial features, for example to distinguish between curve segments from the eyebrow and the upper lip, which have similar transitions in intensities between inner and outer side. In our work, the topology of the face CEM is modeled using shape contexts. In the original shape context approach [7], a shape is represented by a discrete set of sampled points $P = p_1, p_2, \ldots, p_n$. For each point $p_i \in P$, a coarse histogram h_i is computed to define the local shape context of p_i. To ensure that the local descriptor is sensitive to nearby points, the local histogram is computed in a log-polar space. In our case, a histogram is computed for the center point of the curve segment in the face CEM. When considering such a center point, the sampled points P are the discretized points on the other curve segments in the face CEM. An example of a histogram of relative positions is shown in Figure 2, on the right is plotted the log-polar histogram from the curve segment of the right eyebrow. In practice the circle template covers the entire face.

Assume that p_i and q_j are the center points of the curve segments of two different faces. The shape context approach defines the cost of matching the two curve segments by the following χ^2 test statistic:

$$C(p_i, q_j) = \frac{1}{2} \sum_{k=1}^{K} \frac{[h_i(k) - h_j(k)]^2}{h_i(k) + h_j(k)} \tag{2}$$

where $h_i(k)$ and $h_j(k)$ denote the K-bin normalized histograms of relative positions of p_i and q_j, respectively.

4 Applications with Face CEM

4.1 Face Tracking

The face CEMs in consecutive frames are matched using the technique as described in Section 3. While the orientations of the curve axes are used to make a first binary classification, the costs for matching intensity histograms and matching histograms of relative positions are linearly combined for reasons of simplicity. Experiments show that for a pair m of matching curve segments, the linearly combined cost $D_m = a(B_{in} + B_{out})/2 + bC$, where B_{in} and B_{out} are costs for matching intensity histograms and C is the cost for matching histograms of relative positions, has a maximized recognition rate when $a = 0.83$ and $b = 0.17$.

We look for a one-to-one match for each of the curve segments in the current frame by a minimization of the linearly combined cost D_m. For the construction of motion vectors we use similar techniques as proposed in [9]. A motion vector of a unique curve correspondence pair is defined by the center points of two curve segments.

4.2 Face Recognition

The input face CEM is matched with the face CEMs in a database using the technique described in Section 3. From the pairs of matching curve segments, a global length weighted matching cost E is computed, with $E = \sum_{m=1}^{M} D_m l_m / \sum_{m=1}^{M} l_m$, where M is the number of unique pairs of curve segments and l_m is the average length of the matching curve segment pair m. Then, a face is recognized when the global matching cost reaches a minimum.

4.3 Facial Feature Detection

Facial features such as the eyebrows, the eyes, the nose and the lips, are detected using the face CEMs.

Firstly, curve segments from the facial features of interest have been modeled by a training proces on a database, consisting of 100 faces, considering small variations in pose, variations in lighting conditions and variations in facial expressions. The curve segment model of a facial feature consists of an orientation of the axis of the curve segment, two intensity histograms from the inner and outer side of the curve segment and a log-polar histogram describing the relative positions in the face CEM.

Secondly, the curve segment models are matched with the input face CEM. The coefficients of the curve segments make a first binary classification. For the lips and the eyebrows the alignment of the curve segment axes are vertical, while for the nose the alignment of the curve segment axis is horizontal. Next, the histograms of intensities and the histograms of relative positions are compared. E.g. for the upper curve segment of a left eyebrow we expect a transition in intensity from bright to dark and with a relative position in the left upper corner in the face CEM. In this manner, we classify the curve segments from the left and right eyebrow, the left and right eye, the left and right side of the nasal bone and the upper and the lower lip.

5 Results

Our system applications as described in Section 4 are evaluated on different face databases, namely the Georgia Tech Face Database [10], the Database of Faces [11], the Bern University Face Database [12], the AR Face Database [13], the Yale University Face Database [14] and the BioID Database [15] with ground truth marker points.

To show the real-time properties of our system, we give in Table 1 an overview of the computational performance of our face analysis system. The system is implemented in C++ as a webcam application, running on a 2.80 GHz processor, 4.00 GB RAM and 64-bit operating system. We average computational time over 1000 frames for the face detection, the CEM construction and the CEM matching, considering different frame and face sizes.

The CEM matching is evaluated by the application of face recognition as described in section 4.2. The face recognition rates are given in Table 3. A distinction is made between the matching technique described in [6], matching using histograms of relative positions, matching using intensity histograms and matching using an optimal linear combination of the latter two. In the results, the correct match is only counted when the best matched face from a model is the correct person. We test face recognition under controlled condition and size variation, under varying lighting condition, under varying facial expression and under varying pose. As we expect, the matching of histograms with relative positions is more sensitive to varying pose and varying facial expressions, while the matching of histograms is more sensitive to varying lighting conditions. The average face recognition for the linearly combined cost is 95.51%. When compared to the results described for the matching technique in [6], the average face recognition rate increases with 2.21%. Furthermore, we gain the advantage of facial feature detection, which involves facial action recognition, such as recognition of facial expressions, talking and head movement.

The facial feature detection as described in Section 4.3 is evaluated on the test databases by veryfying whether or not the curve segment models classify the correct curve segments in the faces. We classify the curve segments from the left and the right eyebrow, the left and the right eye, the left and the right side of the nasal bone and the upper and the lower lip, as shown in Figures 3 (a) and (b). The results for facial feature detection, as presented in Table 2, show that the developed system can detect on

Table 1. Computational performance: Average computational time for face detection, CEM construction and CEM matching, considering different frame and face sizes

Frame size	320x240	640x480	1280x960
Face detection (ms)	48.85	52.27	102.93
Face size	100x100	200x200	400x400
CEM construction (ms)	16.35	37.88	61.38
CEM matching (ms)	9.16	26.36	43.38
Total (ms)	74.36	116.51	207.69
fps	13.45	8.58	4.81

Table 2. Results for facial feature detection: The detection rates for the left and the right eyebrow, the left and the right eye, the left and the right side of the nasal bone and the upper and lower lip

(%)	YFD	BERN	AR	GTFD	ATT	BioID
Left eyebrow	100.00	90.00	87.91	96.00	92.50	93.12
Right eyebrow	100.00	80.00	90.47	94.00	90.00	94.65
Left eye	93.33	90.00	86.20	86.00	85.00	91.39
Right eye	93.33	93.33	87.91	88.00	87.50	90.94
Left nose	100.00	90.00	88.03	96.00	92.50	94.10
Right nose	100.00	96.67	87.18	90.00	95.00	92.94
Upper lip	93.33	90.00	89.20	86.00	97.50	95.51
Lower lip	100.00	96.67	90.60	86.00	92.50	93.77
Average	**97.50**	**90.46**	**88.44**	**90.25**	**91.56**	**93.30**

Table 3. Results for face recognition: The second, third, fourth and fifth columns show the recognition rates for the matching technique described in [6], matching using histograms of relative positions, matching using intensity histograms and matching using an optimal linear combination of both, respectively

(%)	Old	Hist. of	Intensity	Lin.
	[6]	rel. pos.	hist.	comb.
Controlled condition				
GTFD [10]	98.00	82.00	98.00	100.00
ATT [11]	100.00	77.50	95.00	98.50
YFD [14]	NA	86.66	100.00	100.00
BERN [12]	100.00	93.33	96.67	100.00
AR [13]	98.00	78.65	96.02	98.44
Varying pose				
BERN Right	93.34	70.33	87.00	91.33
BERN Left	86.67	70.67	88.33	91.67
BERN Up	86.67	72.00	86.67	90.33
BERN Down	73.34	69.00	80.33	83.67
Size variation				
AR with size variation	90.21	77.83	89.20	94.35
Varying lighting condition				
AR with left light on	96.34	73.53	91.89	96.60
AR with right light on	94.84	73.43	89.16	95.63
AR with both lights on	92.10	75.15	90.63	96.38
Varying facial expression				
AR with smiling expr.	96.53	71.13	93.70	97.13
AR with angry expr.	97.30	71.59	94.14	97.46
AR with screaming expr.	96.21	72.08	91.40	96.65
Average	**93.30**	**76.18**	**91.76**	**95.51**

average the individual facial features successfully in 91.92% cases, when applied to the test databases. As a comparison, our average facial feature detection rate for the BioID Database is 93.30%, which is comparable with the facial feature detection rate by applying the Viola and Jones face detector [8]. The Viola and Jones face detector finds 95% of facial feature points within 20% of the inter-ocular separation on the BIOID Database [16]. However, experiments show that our approach has the advantages of low computational cost and large invariance for variations in facial expressions.

The accuracy in position of the facial features with the curve segments is determined by the accuracy in position of the edges delivered by the canny edge detector and the fitting cost allowed during segmentation. We compute this accuracy on the BioID Database, by comparing the available ground truth marker points with the locations of the facial features detected. In Figure 3 (a), the ground truth marker points are indicated with red crosses. The criteria for the accuracy is the distance of the points on the curved segments closest to the ground truth markers. When using the mean euclidean distance as the basic error measurement, the mean detection error is 6.53 pixels with a standard deviation of 1.15 pixels, which is 2.18% in terms of the size of the face. We compare to the work of Ding et. al. [17], in which they compute the average accuracy on the AR Face Database and the XM2VT Face Database. The mean detection error for SubAdaBoost is 9.0 pixels with a standard deviation of 1.3 pixels, which is 2.8% in terms of the size of the face.

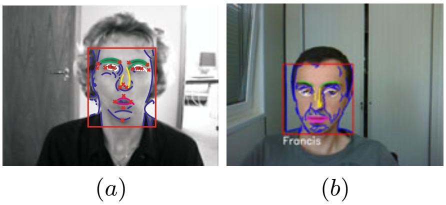


Fig. 3. Detection of facial features: Image (a) shows the curve segments of the eyebrows, the eyes, the nose and the lips detected in a face from the BioID Database [15]. The ground truth marker points are indicated with red crosses. Image (b) shows the facial feature detection in a face from a webcam video sequence.

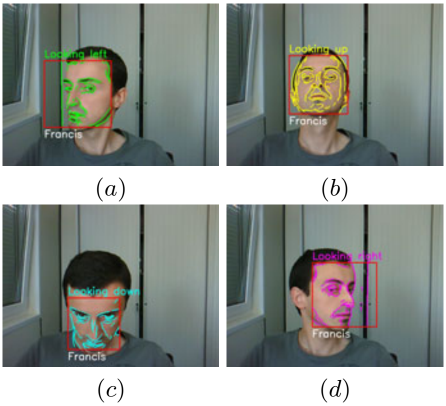


Fig. 4. Results for face tracking: Image (a), (b), (c) and (d) show a face from a webcam video sequence, which is looking left, right, up and down, respectively.

As a result for face tracking, Figures 4 (a), (b), (c) and (d) show the tracking of a face, which is looking left, right, up and down, respectively. From the motion vectors of the matching parabola segment pairs, an average motion vector makes an estimation about the head movement. In future work, facial action recognition, such as facial expression recognition, head movement recognition and talking recognition will be further explored.

6 Conclusion

This paper proposes an alternative way to the analysis of faces. Our method models faces with Curve Edge Maps, which is a natural description for the many edges in a face and correspond to physically meaningful features, such as the eyebrows, the eyes, the nose and the lips. We presented a novel matching technique, which uses the orientation of the axes of the curve segments, intensity histograms and histograms of relative positions. We demonstrated that the CEM approach is useful for many face analysis tasks, such as face tracking, face recognition and facial feature detection. Applied to different databases, good performance for these applications is achieved, which is comparable with or better than the results in literature. However, our method provides simplicity, real-time performance and extensibility to the different aspects of face analysis.

References

1. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Training Models of Shape from Sets of Examples. In: Proc. British Machine Vision Conference, pp. 9–18 (1992)
2. 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)

3. Liu, X.: Generic face alignment using boosted appearance model. In: Proc. IEEE Computer Vision and Pattern Recognition, pp. 1079–1088 (2007)
4. Liu, X.: Discriminative Face Alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(11), 1941–1954 (2009)
5. Liang, L., Xiao, R., Wen, F., Sun, J.: Face alignment via component-based discriminative search. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 72–85. Springer, Heidelberg (2008)
6. Deboeverie, F., Veelaert, P., Teelen, K., Philips, W.: Face Recognition Using Parabola Edge Map. In: Blanc-Talon, J., Bourennane, S., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2008. LNCS, vol. 5259, pp. 994–1005. Springer, Heidelberg (2008)
7. Belongie, S., Malik, J., Puzicha, J.: Shape Matching and Object Recognition Using Shape Contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 509–522 (2001)
8. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. of Computer Vision and Pattern Recognition, pp. 511–518 (2001)
9. Deboeverie, F., Teelen, K., Veelaert, P., Philips, W.: Vehicle Tracking Using Geometric Features. In: Blanc-Talon, J., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2009. LNCS, vol. 5807, pp. 506–515. Springer, Heidelberg (2009)
10. Georgia Institute of Technology, Georgia Tech Face Database, http://www.anefian.com/face_reco.htm
11. AT&T Laboratories, Cambridge, The Database of Faces, http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html
12. University of Bern, Bern, Switzerland, Bern University Face Database, ftp://iamftp.unibe.ch/pub/ImagesFaceImages/
13. Martinez, A.M., Benavente, R.: The AR Face Database, CVC Technical Report #24 (1998)
14. University of Yale, Bern, Switzerland, Yale University Face Database, http://cvc.yale.edu/projects/yalefaces/yalefaces.html
15. Jesorsky, O., Kirchberg, K.J., Frischholz, R.W.: Robust face detection using the hausdorff distance. In: Bigun, J., Smeraldi, F. (eds.) AVBPA 2001. LNCS, vol. 2091, pp. 90–95. Springer, Heidelberg (2001)
16. Cristinacce, D., Cootes, T.F.: Automatic Feature Localisation with Constrained Local Models. Pattern Recognition 41(10), 3054–3067 (2008)
17. Ding, L., Martinez, A.M.: Features versus Context: An Approach for Precise and Detailed Detection and Delineation of Faces and Facial Features. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(11), 2022–2038 (2010)
18. Deboeverie, F., Teelen, K., Veelaert, P., Philips, W.: Adaptive Constructive Polynomial Fitting. In: Blanc-Talon, J., Bone, D., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2010, Part I. LNCS, vol. 6474, pp. 173–184. Springer, Heidelberg (2010)
19. Canny, J.: A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 8(6), 679–698 (1986)
20. Karande, K.J., Talbar, S.N.: Independent Component Analysis of Edge Information for Face Recognition. International Journal of Image Processing 3(3), 120–130 (2009)
21. Veelaert, P., Teelen, K.: Fast polynomial segmentation of digitized curves. In: Kuba, A., Nyúl, L.G., Palágyi, K. (eds.) DGCI 2006. LNCS, vol. 4245, pp. 482–493. Springer, Heidelberg (2006)
22. Cha, S.-H., Srihari, S.N.: On measuring the distance between histograms. Pattern Recognition 35(6), 1355–1370 (2002)

Statistical Patch-Based Observation for Single Object Tracking

Mohd Asyraf Zulkifley and Bill Moran

Department of Electrical and Electronic Engineering
The University of Melbourne
Victoria 3010 Australia
m.zulkifley@student.unimelb.edu.au, wmoran@unimelb.edu.au

Abstract. Statistical patch-based observation (SPBO) is built specifically for obtaining good tracking observation in robust environment. In video analytics applications, the problems of blurring, moderate deformation, low ambient illumination, homogenous texture and illumination change are normally encountered as the foreground objects move. We approach the problems by fusing both feature and template based methods. While we believe that feature based matchings are more distinctive, we consider that object matching is best achieved by means of a collection of points as in template based detectors. Our algorithm starts by building comparison vectors at each detected point of interest between consecutive frames. The vectors are matched to build possible patches based on their respective coordination. Patch matching is done statistically by modelling the histograms of patches as Poisson distributions for both RGB and HSV colour models. Then, maximum likelihood is applied for position smoothing while a Bayesian approach is applied for size smoothing. Our algorithm performs better than SIFT and SURF detectors in a majority of the cases especially in complex video scenes.

Keywords: Neyman-Pearson, Tracking observation, Poisson modelling, Maximum likelihood.

1 Introduction

Object detection algorithms have evolved from the implementation of simple edge matching to complex feature matching. In multiple objects tracking systems, good matching technique is important for distinguishing objects between consecutive frames. Robust matching algorithms provide better measurement inputs for updating tracking algorithms. Early feature based object matching algorithms are built on edge information such as Canny [1], Sobel [2] and Roberts [3] edge operators. The detected edges are compared to find similarity traits between objects. In order to improve the detection accuracy, corner detectors such as the algorithms of Harris [4], and Shi and Tomasi [5] are used for finding corners which serve as the points of interest. However, the resulting points are not very distinctive and perform poorly under illumination change. In 1999, Lowe introduced the Scale Invariant Feature Transform (SIFT), [**?**], that generates robust features that can cater for problems of rotation, scaling and moderate illumination

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 119–129, 2011.

change. Ke and Suthankar [6] improved SIFT's feature distinctiveness by applying principal components analysis. Later, Burghouts and Geusebroek [7] fused SIFT with their colour invariance algorithm [8] to achieve robustness to the illumination change. Performance evaluation of SIFT and its variants are explained in detail by Mikolajczyk and Schmid [9]. They also introduced their own feature descriptor's algorithm, GLoH that enhanced SIFT by integrating more spatial properties during histogram accumulation. In 2006, Bay et al. [10] introduced Speeded Up Robust Features (SURF) which has similar detection performance to SIFT but with a much lower computational burden. However, all these algorithms are based on fixed size kernel matching in which each pixel descriptor is obtained based on certain pixel distance from the anchor pixel. For example, SIFT uses fix kernel size of 16×16 pixels. Recognition performance degrades when the objects are blurred, moderately deformed or possess homogenous texture. Another approach is use template based matching, which requires several templates to be matched by correlation over the search regions. This approach allows a collection of pixels to be compared and matched as a group. This idea is basically the antithesis of a content-based image retrieval (CBIR) approach and provides better robustness in detecting moderately deformed objects. However, its disadvantages include the fact that the search area is usually big and the chosen templates need to be able to cater all possible transformations of the object between frames.

In order to overcome the weaknesses of both approaches, we have fused both feature and template based methods together. The aim of this paper is to find the matches of foreground objects between consecutive video frames especially in complex scenes for updating tracking algorithms observation. We assume that the size of the objects between the consecutive frames does not change too much. The advantage of using fusion approach is it manages to obtain good detection even for the case of object deformation as normally occurs during human movement. Use of the template-based approach alone results in algorithms that have a diminished distinctiveness property because most objects are nonrigid; while for the implementation of feature based detector alone, it is bound to fail as some points are covered or hidden in the later frames. One major difference between image processing and video analytics is the sharpness of the captured image. Usually, in the former one assumes that the image captured has sharp boundaries, while this is not necessarily the case for video. Most feature based algorithms are not able to obtain a match in this context. They also perform poorly in matching objects under low ambient illumination.

Our algorithm is intended to increase robustness of detecting tracking observation in situations where there is a blurring effect, illumination change, low illumination surroundings, homogenous texture and moderate deformed objects. The main idea is to match the same points of interest between frames and use this to develop possible patches for object matching. However, this is performed only to selected points of interest so that the calculation burden is reduced compared to a simple template based approach. To improve detection accuracy, position smoothing is performed by aligning the bounding box to the object centroid. Lastly, size adjustment is performed to cater for small change in object size. Any object will become bigger as it moves closer to the camera and smaller as it recedes. In addition, deformation usually forces a new bounding box size. Finally, the output patch is fed into any filter-based tracker as the measurement input.

2 Statistical Patch-Based Observation (SPBO)

SPBO is built specifically for video applications that require object matching between consecutive frames. This algorithm has a moderate distinctiveness property yet better recognition accuracy for most applications. The challenge lies in recognizing objects under illumination change, blurring effects, moderate deformation and in handling objects with homogenous texture. The main components of the algorithm are 1) generate possible patches, 2) undertake patch matching, 3) perform position smoothing and 4) perform size smoothing.

2.1 Generate Possible Patches

Point of interest are first used to find the locations to generate vector descriptors. These descriptors are matched between frames for building possible patches. The original location of the object or the original patch in the first frame is initialized by the user. The importance of this patch is that it serves as the reference for building the statistical data used in matching and smoothing procedures; in particular the reference histograms. Moreover, the size of the first frame patch is an indicator of the size of the original object. Let P_w and P_h denote the width and height of the user defined bounding box, while (x, y) represents the coordinate location and t is the time of the frame. Both, the first frame ($F^{x,y,t}_{r,g,b}$) and the second frame ($F^{x,y,t+1}_{r,g,b}$) are transformed to greyscale space ($F^{x,y,t}_{I}$, $F^{x,y,t+1}_{I}$). Corner detectors as defined by Shi and Tomasi [5] are applied to find the possible points of interest. The algorithm was chosen because of its ability to generate the points even during low ambient illumination and for low textured objects. For the first frame, the points are generated inside the predefined patch only, while for the second frame, the points are generated for the whole image. A vector descriptor, **V** is used for matching the points of interest between frames. Possible bounding boxes are generated at each corner where the vectors are matched. There will be 3 sets of vectors for each point of interest, ($\mathbf{V}^{x,y,t}_R$, $\mathbf{V}^{x,y,t}_G$, $\mathbf{V}^{x,y,t}_B$). Let i denote the channel type and define

$$\mathbf{V}^{x,y,t}_i = \{F^{x-1,y,t}_i - F^{x,y,t}_i, F^{x,y-1,t}_i - F^{x,y,t}_i, F^{x+1,y,t}_i - F^{x,y,t}_i, F^{x,y+1,t}_i - F^{x,y,t}_i\} \quad (1)$$

The vectors are generated by finding the colour difference between the anchor pixel and its selected neighbourhood pixels as shown in Figure 1. Then the vector components are

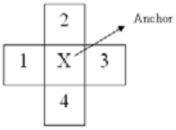


Fig. 1. Neighbourhoods pattern used for vectors generation

sorted from the lowest to the highest value. Each vector from an initial frame is compared with each vector in the next frame. The decision rule (dr) for vectors matching is shown in equation 2 where the differences between each vector component are summed up and the final value is obtained by combining all 3 channels differences. Then it is compared with a predefined threshold, γ_1 which was found by experiment to be optimal in the range of 11 to 13. Let $L_1^{x,y,t}$ denotes the label which takes the value 1 when the vectors are matched and 0 for unmatched vectors.

$$dr = \sum_{i=R}^{B} \left| \mathbf{V}_i^{x,y,t} - \mathbf{V}_i^{x,y,t-1} \right| \tag{2}$$

$$L_1^{x,y,t} = \begin{cases} 1 \text{ if } dr < \gamma_1 \\ 0 \text{ if } dr \geq \gamma_1 \end{cases} \tag{3}$$

All of the matched vectors are candidate for locations at which patches are built. Patches for the second frame are generated around the location of the matched vector in the first frame with respect to the original bounding box. Figure 2 shows an example of how the bounding box is generated. Initially, the size of the object is assumed to remain constant between frames.

A subsequent test for distinguishing overlapping patches is performed after all patches have been assigned location and size. This is done in order to reduce the calculation burden by reducing the number of patches. Patch smoothing is performed if the overlapping area is more than 70% of the original patch size.

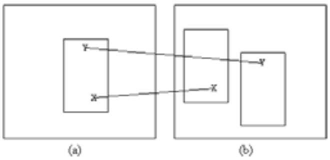


Fig. 2. Examples of constructing the new patches between the frames. The bounding boxes are aligned with respect to the matched vectors in the first frame. (a) First frame (b) Second frame.

2.2 Patch matching

Patch matching is performed to find the patch where the object most likely resides. The match is done by comparing the histograms of the first and second frame patches. Two colour models are considered: RGB and HSV colour spaces. For the case of no illumination change, use of RGB colour space gives a better histogram comparison. When illumination change occurs, the hue channel from the HSV colour model gives better comparison since the hue channel remain invariant but the distinctive feature are degraded. For RGB colour space, a 3-dimensional histogram is built for each patch while a 1-dimensional histogram is built for the hue channel. All histogram matching is done by modelling the relationship between two histograms as a Poisson distribution as in 4 and 5. We chose Poisson distribution as it gives good probability density function for histogram matching which later will be integrated into Bayesian-based decision. Let

N_b be the number of histogram bins in 1-dimensional, n_i and m_i denotes the i^{th} bin value of the first and second frame histograms respectively.

For the 1-dimensional histogram:

$$P(n_{(i)}, m_{(i)}) = \prod_{i=1}^{N_b} \left(\frac{\exp^{-n_{(i)}} n_i^{m_{(i)}}}{m_{(i)}!} \right), \tag{4}$$

and for the 3-dimensional histogram:

$$P(n_{(i,j,k)}, m_{(i,j,k)}) = \prod_{i=1}^{N_b} \prod_{j=1}^{N_b} \prod_{k=1}^{N_b} \left(\frac{\exp^{-n_{(i,j,k)}} n_{(i,j,k)}^{m_{(i,j,k)}}}{m_{(i,j,k)}!} \right) \tag{5}$$

A maximum likelihood approach is used to find the matched patch for both colour models. The likelihoods are modelled by equations 4 and 5 where $\hat{\beta}$ denotes the matched patch and **x** represents the observation.

$$P(\mathbf{x}|\beta) = \begin{cases} P(n_{(i)}, m_{(i)}) \text{ for HSV colour model} \\ P(n_{(i,j,k)}, m_{(i,j,k)}) \text{ for RGB colour model} \end{cases} \tag{6}$$

$$\hat{\beta} = \underset{\forall \beta}{\text{argmax}}\, P(\mathbf{x}|\beta) \tag{7}$$

There are two candidates for the most likely patch. The decision to choose hue over the RGB colour model is decided by using a Neyman-Pearson hypothesis testing [11]. Let $P(\mathbf{x}; \mathcal{H}_0) = P(\hat{\beta}_{\text{RGB}})$, $P(\mathbf{x}; \mathcal{H}_1) = P(\hat{\beta}_{\text{hue}})$ and λ_1 represent the threshold for the Neyman-Pearson hypothesis testing. If the test favours $\mathcal{H}_0$, then an indicator, ϵ is initialized as 1, while if $\mathcal{H}_1$ is chosen, ϵ is equal to 0. The parameter ϵ is the indicator for deciding which colour space is used for the position and size smoothing. The resultant patch (β_{fin_1}) from the test will be the final matched patch.

$$\text{NP}_1 = \frac{P(\mathbf{x}; \mathcal{H}_1)}{P(\mathbf{x}; \mathcal{H}_0)} = \frac{P(\hat{\beta}_{\text{hue}})}{P(\hat{\beta}_{\text{RGB}})} > \lambda_1 \tag{8}$$

$$\beta_{\text{fin}_1} = \begin{cases} \hat{\beta}_{\text{RGB}} \text{ if } P(\hat{\beta}_{\text{hue}}) < \lambda_1 P(\hat{\beta}_{\text{RGB}}) \\ \hat{\beta}_{\text{hue}} \text{ if } P(\hat{\beta}_{\text{hue}}) \geq \lambda_1 P(\hat{\beta}_{\text{RGB}}) \end{cases} \tag{9}$$

2.3 Position Smoothing

Position smoothing is used to adjust the centroid of the patch to accurately align with the centroid of the object. Sometimes, the calculated patch is slightly misaligned with the original object which is prevalent during illumination change or in low ambient illumination. There are two techniques which the patch position is adjusted depending on the value of ϵ with the RGB histogram applied for $\epsilon = 1$ and the hue histogram for $\epsilon = 0$. Firstly, the step size used for adjusting the patch translation is determined, $\delta = \alpha(\min(P_w, P_h))$. Let α denote a weight factor which is found experimentally to be optimal within $[0, 0.5]$ based on the assumption that the object size does not change

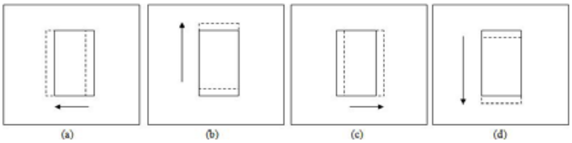


Fig. 3. Patches coordination for location smoothing (a) Left side translation (b) Upward translation (c) Right side translation (d) Downward translation

abruptly between two consecutive frames. Four new candidate patches are created for adjusting the patch position as shown in Figure 3, representing translations in four directions of the pivot patch ($\beta_{\text{fin}_1}^{\text{old}_0}$): leftward ($\beta_{\text{fin}_1}^{\text{new}_1}$), upward ($\beta_{\text{fin}_1}^{\text{new}_2}$), rightward ($\beta_{\text{fin}_1}^{\text{new}_3}$) and downward ($\beta_{\text{fin}_1}^{\text{new}_4}$). The histograms of each of the five patches including the original position patch are obtained, and again maximum likelihood is used to find the new location. No histogram normalization is needed in this subsection as both original and candidate patches are of the same size. Likelihood is derived from the relationship between the first and second frame histograms as in equations 4 and 5. Let $\hat{\beta}_{\text{fin}_2}$ denotes the output of the position smoothing.

$$P(\mathbf{x}|\beta_{\text{fin}_1}) = \begin{cases} P(n_{(i)}, m_{(i)}) \text{ if } \epsilon = 0 \\ P(n_{(i,j,k)}, m_{(i,j,k)}) \text{ if } \epsilon = 1 \end{cases} \tag{10}$$

$$\beta_{\text{fin}_2} = \underset{\forall \beta_{\text{fin}_1}}{\text{argmax}}\, P(\mathbf{x}|\beta_{\text{fin}_1}) \tag{11}$$

For each iteration, the pivot position is reinitialized by letting $\beta_{\text{fin}_1}^{\text{old}_0} = \beta_{\text{fin}_2}$, so that 4 new translated patches for the next iteration are built around β_{fin_2}. The algorithm is iterated until the estimated patch position remain the same as shown by the decision rule L_2.

$$L_2 = \begin{cases} \beta_{\text{fin}_2} = \beta_{\text{fin}_1}^{\text{old}_0} \text{ stop the iteration} \\ \beta_{\text{fin}_2} \neq \beta_{\text{fin}_1}^{\text{old}_0} \text{ continue the iteration} \end{cases} \tag{12}$$

2.4 Size Smoothing

This section focuses on adjusting the size of the patch so that it provides a good fit to the object. Generally, the image of the object becomes bigger as it moves closer to the camera and smaller as it moves away. However, the size increment and decrement between the frames will not be very large. We limit the scale change for size smoothing by at most a factor of $\frac{1}{2}$. Eight new patches with different sizes are used for the size smoothing test. The same δ used in position smoothing is used to adjust the patch size. Four shrinkage and four expansion pattern patches are obtained either by subtracting from or adding to one of the patch corners by a step size value. Figure 4 shows a shrunk patch, while an expanded patch is shown in Figure 6. $\hat{\beta}_{\text{fin}_2}$ is the pivot point for creating all the new patches. A Bayesian approach is used to decide the final patch size ($\hat{\beta}_{\text{fin}_3}$) from among the nine patches including the original patch ($\hat{\beta}_{\text{fin}_2}$).

$$P(\beta_{\text{fin}_2}|\mathbf{x}) \propto P(\mathbf{x}|\beta_{\text{fin}_2})P(\beta_{\text{fin}_2}) \tag{13}$$

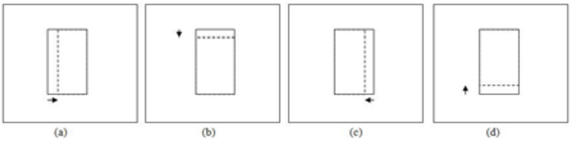


Fig. 4. New shrinking patches pattern (a) Left side shrinkage (b) Upper side shrinkage (c) Right side shrinkage (d) Lower side shrinkage

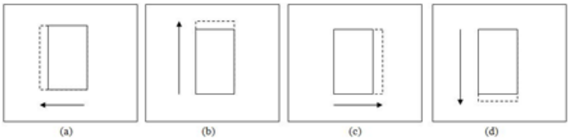


Fig. 5. New expanding patches pattern (a) Left side expansion (b) Upper side expansion (c) Right side expansion (d) Lower side expansion

The ϵ value determines what type of histogram is built. If $\epsilon = 0$, a 1-dimensional hue histogram is used while for $\epsilon = 1$, a 3-dimensional RGB histogram is applied. Before any comparison is performed, the histogram size needs first to be normalized. Let S_1 and S_2 denote the number of pixels in the patches in the first and second frames respectively. Each histogram bin value, H is adjusted by the ratio of sizes of S_2 and S_1, $H^{\text{new}} = \left(\frac{S_2}{S_1}\right) H^{\text{old}}$. Once again, the histogram relationship between the first and second frame patches are modelled by a Poisson distribution.

$$P(\mathbf{x}|\beta_{\text{fin}_2}) = \begin{cases} P(n_{(i)}, m_{(i)}) \text{ if } \epsilon = 0 \\ P(n_{(i,j,k)}, m_{(i,j,k)}) \text{ if } \epsilon = 1 \end{cases} \tag{14}$$

Two sets of prior probabilities are used. These are dependent on whether the size of the detected object inclines towards expansion $(P(\beta_{\text{fin}_2}^{\text{big}_p}))$ or shrinkage $(P(\beta_{\text{fin}_2}^{\text{small}_p}))$ where p^{th} denotes the patch under consideration. The selection of a suitable prior is very important as the shrinkage likelihoods are usually large even when the object expands. Thus we apply lower prior probabilities to shrinkage candidates if the size is increasing. In order to determine which set of the priors to be used, again a Neyman-Pearson hypothesis test is implemented where $\mathcal{H}_0$ and $\mathcal{H}_1$ represent the expansion and shrinkage hypotheses. Only eight candidate patches are used (4 shrinkage patches + 4 expansion patches) for this test where the same Poisson distribution is used as in equation 14. The maximum probability among the expansion patches represents the $\mathcal{H}_0$ probability while the maximum probability among the shrinkage patches represents the $\mathcal{H}_1$ probability.

$$P(\mathbf{x}; \mathcal{H}_0) = \max_{\forall \text{big}} P(\mathbf{x}|\beta_{\text{fin}_2}) \tag{15}$$

$$P(\mathbf{x}; \mathcal{H}_1) = \max_{\forall \text{small}} P(\mathbf{x}|\beta_{\text{fin}_2}) \tag{16}$$

Let λ_2 be the threshold for the Neyman-Pearson test.

$$\mathrm{NP}_2 = \frac{P(\mathbf{x};\mathcal{H}_1)}{P(\mathbf{x};\mathcal{H}_0)} > \lambda_2 \tag{17}$$

$$P(\beta_{\mathrm{fin}_2}) = \begin{cases} P(\beta_{\mathrm{fin}_2}^{\mathrm{big}}) \text{ if } \mathcal{H}_0 \text{ is true} \\ P(\beta_{\mathrm{fin}_2}^{\mathrm{small}}) \text{ if } \mathcal{H}_1 \text{ is true} \end{cases} \tag{18}$$

After the prior is obtained, the posterior probabilty $P_{i^{th}}(\beta_{\mathrm{fin}_2}|\mathbf{x})$ of each of the nine patches is calculated. Each side of the bounding box can be expand or shrink independently based on the L_3 decision rule. Each side size is altered depending on whether the new posteriors exceed the original size posterior. $L_3 = 1$ indicates that the size is updated while $L_3 = 0$ indicates that the size change remain constant and i^{th} takes value from 1 to 8.

$$L_3 = \begin{cases} 0 \text{ if } P_{i^{th}}(\beta_{\mathrm{fin}_2}|\mathbf{x}) \leq P_{0^{th}}(\beta_{\mathrm{fin}_2}|\mathbf{x}) \\ 1 \text{ if } P_{i^{th}}(\beta_{\mathrm{fin}_2}|\mathbf{x}) > P_{0^{th}}(\beta_{\mathrm{fin}_2}|\mathbf{x}) \end{cases} \tag{19}$$

Figure 6 shows two examples of how the size of the object is updated. The iteration is terminated if no size change is detected.

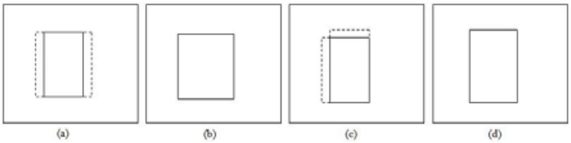


Fig. 6. Example of patch expansion (a)(c) Original patch (b) Result if the right and left side expansion are true (d) Result if the left and upper side expansion are true

3 Results and Discussion

SPBO has been tested on several video sequences that contain moving objects in various frame sizes. SIFT and SURF are chosen as the benchmarks for performance comparison. The implementation of SURF algorithm is based on OpenSURF by Evans [12] while SIFT is applied based on the original Lowe algorithm [**?**]. The performance is measured by calculating the Euclidean distance, $\mathcal{D}$ between the centroid (Ω_{sim}) of the simulation result and the manually determined ground truth centroid (Ω_{truth}) of the detected object.

The 4 corners of SPBO bounding box are used as the reference points for the centroid calculation. While the centroid for SIFT and SURF is generated by constructing a bounding box which uses the extreme points in 4 directions as shown in Figure 7. Then, the generated bounding box corners are used for the centroid calculation. Table 1 shows the distance error analysis among the methods. SPBO performance is the best with 43.3% of the detected object centroid have less than 10 pixels distance from the

Table 1. Comparison of the centroid distance among SPBO, SIFT and SURF

Method	Distance error (pixel)										
	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	>99
SPBO	52	30	19	13	3	0	1	0	0	0	2
SIFT	40	19	9	3	6	8	2	0	2	2	30
SURF	30	15	8	6	5	3	3	1	0	2	47

ground truth centroid. SIFT based recognition manages 33.3% detection with less than 10 pixels distance error while the worst is SURF with just 25% detection. The distance error relative to the patch size is given by this equation, $err = \frac{\mathcal{D}}{\max(P_w^{\text{new}}, P_h^{\text{new}})} \times 100\%$ where SPBO error rate is just 7% while SIFT is 27.1% and SURF is 42.6%. Figure 8 shows some of the results of applying SPBO in various scenes and situations. The images in 8(a) to 8(i) show the results of applying each method to 3 sets of video sequences. All methods obtain good recognition for the first set of video sequences [8(a)-8(c)]. For the second set of images [8(d)-8(f)], only SPBO recognizes the person under an illumination change. While, SIFT gives wrong matching and SURF provides no matching at all. In the third set, only SPBO 8(d) is able to recognize the object under blurring noise because of the fast movement of the ball. On the other hand, both SIFT and SURF did not detect any matched points. Examples of successful implementation of SPBO in various scenes and situations are given in the images 8(j) to 8(o) which include the problems of size change, blurring effect, homogenous texture, deformed object and illumination change.

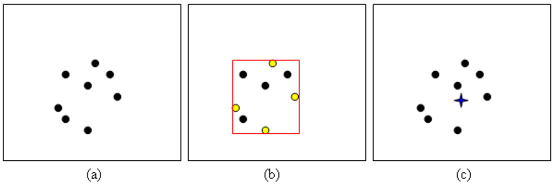


Fig. 7. Generating centroid for the SIFT and SURF algorithms (a) Matched points of interest (b) Constructing a bounding box (c) Centroid location is indicated by the blue star

4 Conclusion

In conclusion, we have shown that SPBO works well for recognizing the objects through video sequences. The problems of image sharpness, moderate deformation, illumination change, blurring, small size change and homogenous texture are solved by fusing both feature and template based approaches. The feature detector is obtained to generate the possible bounding boxes while the matching is done by taking a collection of pixels instead of a single point. This method is suitable for application in a system that requires object recognition in complex scenes.

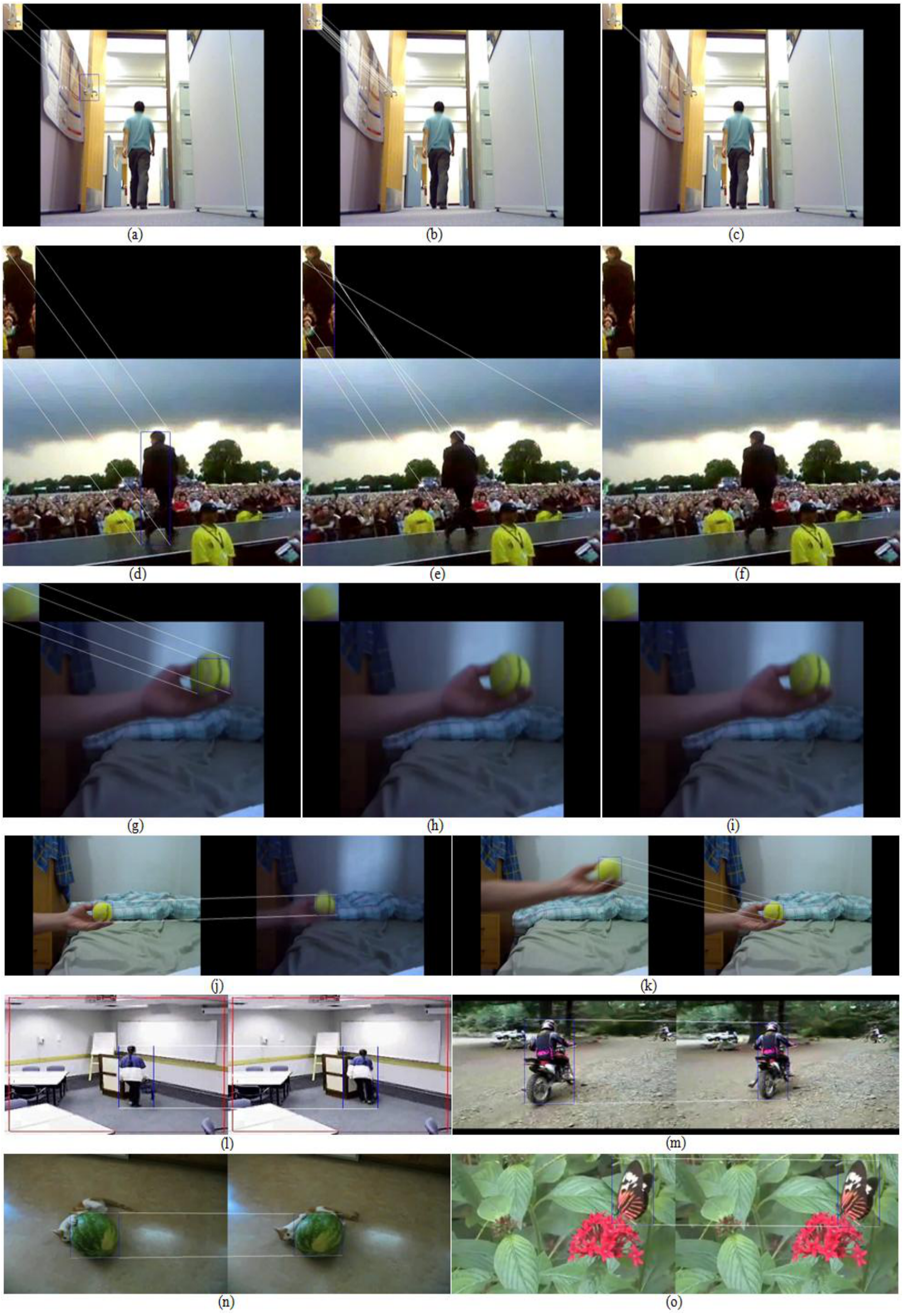


Fig. 8. Results of applying SPBO, (a)-(f): Detection comparison among SPBO [a][d], SIFT [b][e] and SURF [c][f], (g)-(n): Samples of SPBO result (left image is the 1^{st} frame, right image is the 2^{nd} frame)

References

1. Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 679–698 (1986)
2. Duda, R.O., Hart, P.E.: In: Sobel, I., Feldman, G. (eds.) A 3x3 Isotropic Gradient Operator for Image Processing, pp. 271–272 (1973)
3. Roberts, L.G.: Machine perception of three-dimensional solids. PhD thesis, Dept. of Electrical Engineering, Massachusetts Institute of Technology (1963)
4. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proceedings of the 4th Alvey Vision Conference, pp. 147–151 (1988)
5. Shi, J., Tomasi, C.: Good features to track. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 593–600 (1994)
6. Ke, Y., Suthankar, R.: Pca-sift: a more distinctive representation for local image descriptors. In: IEEE Computer Society Conference Computer Vision and Pattern Recognition, vol. 2, pp. 506–513 (2004)
7. Burghouts, G.J., Geusebroek, J.M.: Performance evaluation of local colour invariants. Computer Vision and Image Understanding 113, 48–62 (2008)
8. Geusebroek, J.M., Smeulders, A.W.M., Boomgaarda, R.V.D.: Color invariance. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 1338–1350 (2001)
9. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 1615–1630 (2005)
10. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer Vision and Image Understanding 110, 346–359 (2008)
11. Kay, S.M.: Fundamentals of Statistical Signal Processing, Detection Theory, vol. 2. Prentice Hall, Englewood Cliffs (1998)
12. Evans, C.: Notes on the opensurf library. Technical Report CSTR-09-001, University of Bristol (January 2009)

Exploiting Depth Information for Indoor-Outdoor Scene Classification

Ignazio Pillai, Riccardo Satta, Giorgio Fumera, and Fabio Roli

Deparment of Electrical and Electronic Engineering,
Univ. of Cagliari Piazza d'Armi, 09123 Cagliari, Italy
{pillai,riccardo.satta,fumera,roli}@diee.unica.it

Abstract. A rapid diffusion of stereoscopic image acquisition devices is expected in the next years. Among the different potential applications that depth information can enable, in this paper we focus on its exploitation as a novel information source in the task of scene classification, and in particular to discriminate between indoor and outdoor images. This issue has not been addressed so far in the literature, probably because the extraction of depth information from two-dimensional images is a computationally demanding task. However, new-generation stereo cameras will allow a very fast computation of depth maps. We experimentally show that depth information alone provides a discriminant capability between indoor and outdoor images close to state-of-the art methods based on colour, edge and texture information, and that it allows to improve their performance, when it is used as an additional information source.

Keywords: scene classification, depth map, indoor-outdoor.

1 Introduction

Scene classification is a challenging topic in computer vision. Discriminating between indoor and outdoor images is a particular instance of scene classification which has been addressed by many authors, since it is often at the root of scene classification taxonomies [8,9,21,24]. Its applications range in various fields, including personal photo tagging, image retrieval [25], colour constancy [3], and robotics [6]. Several methods have been proposed so far to face this problem. Most of them are based on extracting low-level information embedded in the image pixels, both in the spatial domain (like colour moments and histograms) and in the frequency domain (through wavelets and the discrete cosine transform). Some recent works exploit meta-data as another source of information.

We argue that a further information source which can convey some discriminant capability between indoor and outdoor scenes is the depth map of the image, namely a map that associates each pixel of the image with its distance computed with respect to the observer.

To our knowledge, depth information was exploited to improve performance in tasks related to the more general field of scene understanding only in [11]. Here different modules (scene classification, image segmentation, object detection,

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 130–139, 2011.

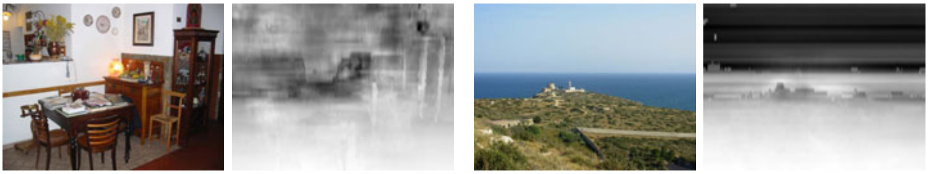


Fig. 1. Example of indoor (left) and outdoor (right) images and corresponding relative depth maps, estimated through [15]. Darker colours correspond to higher depth values.

and depth estimation) were combined to improve the performance of each one, by providing the output of other modules as an additional input. Instead, the usefulness of the depth information in discriminating among different scenes has not been investigated yet, although suggested in [23].

A depth map can be easily obtained from a stereo image pair (i.e., two images of the same scene taken from slightly different viewpoints) while its estimation from a single image is computationally demanding. Stereo acquisition devices are not yet widespread and this is perhaps one of the reasons why the use of depth information for scene classification has not been explored. However, thanks to the growing interest in 3D imaging, many manufacturers have presented (or are planning to present) devices able to take stereo pictures and videos.

Intuitively, depth maps of outdoor scenes are likely to exhibit higher depth values than indoor ones. However, absolute depth values, corresponding to real distances, can not be obtained unless all the parameters of the stereo camera system are exactly known. Moreover, there is always a practical limit to the maximum depth value that can be measured, which generally lies around 30 meters [26]. Still, in any case a *relative* depth map can be computed, whose values are relative ranking of pixels based on their depth. Our intuition is that even relative depth maps embed in their "structure" enough information to discriminate between indoor and outdoor scenes (see Fig. 1).

Based on the above motivations, in this paper we experimentally investigate the usefulness of relative depth maps as a novel information source for indoor-outdoor scene classification. To this aim, we propose three possible feature sets extracted from the relative depth map, based on the analysis of the pixel depth distribution in two publicly available image data sets. We then evaluate the discriminant capability of these feature sets, when they are used as the only information source for a classification algorithm, as well as when they are used as additional information sources to improve existing methods. Since no suitable stereoscopic image data set is available, we carried out experiments on two publicly available corpora of single images, estimating their relative depth maps using a recently proposed method. These depth maps can be seen as an approximation of those obtainable in real application scenarios (relative depth maps estimated by stereo pairs).

Our results show that proposed feature sets exhibit a good discriminant capability for the indoor-outdoor problem. Furthermore, we show that depth information allows to improve the performance of state-of-the art methods, due to its complementarity with pixel-based information. Since these results have been obtained in an unfavourable setting (approximated depth maps), performances in real scenarios should be even better.

In Sect. 2 we survey previous works on indoor-outdoor scene classification. Basic information on depth map computation, including the method used in this work, is provided in Sect. 3. The features we devised based on depth maps are described in Sect. 4. In Sect. 5, we report experimental results. Conclusions and further research directions are summarised in Sect. 6.

2 Related Works on Indoor-Outdoor Classification

Most of the works on indoor-outdoor scene classification rely on low-level features based on colour, like histograms [7,17,18,20,21], and moments [10,16]. Many approaches are also based on textures [2,10,16,17,18,21,22] and/or on edges [7,14,16,21]. In a few works more complex features are also used, for example based on entropy of the pixel values [21], or shape [10]. Features are extracted either from the whole image [21], or from regions obtained by a predefined rectangular block subdivision [2,14,17,18,20] or by image segmentation [7,10]. Other works are based on bags of visual words, e.g. [2]. Recently, some authors proposed to combine pixel-level image information with the meta-data often associated to images, like *EXIF* camera informations [4,12,19] and user-generated tags [12].

In the following we focus on the methods in [17,21], which can be considered representative of the pixel-based works mentioned above, are claimed to attain a high accuracy, and provide as well enough information for their implementation. We also consider two feature sets that have proven to attain a high performance in various scene classification tasks, including the indoor-outdoor problem: the *Gist* descriptor [13] and the *Centrist* descriptor [27]. These methods will be used in the experiments of Sect. 5, and are described in the rest of this section.

The approach of [17] is based on two different feature sets, extracted from 4×4 rectangular image sub-blocks of identical size: colour features (histograms on the LST colour space), and texture features (energy of the sub-bands of a two-level wavelet decomposition). Each sub-block is classified by two SVM classifiers, based respectively on colour and texture features. The sum of the classifier outputs over all sub-blocks is computed separately for each feature set; the two resulting values are then fed to another SVM classifier, which provides the final label. In a subsequent work the authors evaluated the performance improvement obtainable through sky and grass detectors [18]. However, only by using ground truth information they obtained significant improvements in respect to [17]. For this reason, we considered [18] not profitable in a realistic classification scenario, and thus chose to implement [17].

In [21], a two-stage approach as in [17] was proposed. The input image is represented using seven different feature set, based on colour distribution, wavelet decomposition, entropy and edge directions. Each feature vector is fed to a distinct neural network. The outputs of the seven first-stage classifiers are then combined by another neural network, which provides the label of the image.

In [13] an image *Gist* descriptor was proposed, based on a set of holistic properties capable to represent the spatial structure of a scene. Such properties were estimated by means of spectral and coarsely localized information.

The *Centrist* descriptor proposed in [27] is based on the Census Transform, a per-pixel transform originally designed for matching among local patches. Histograms of the transformed image are computed at different scales by exploiting spatial pyramids.

3 Depth Map Estimation

The depth map of a scene from a stereo image pair can be computed using several methods, which exhibit a low computational time, suitable even for real time applications. A comprehensive survey is reported in [26]. The simplest way to compute the depth map is to exploit binocular disparity. Given a plane parallel to the ones where the pair of two-dimensional images lie, the distance of a given point in the scene from that plane can be computed by using triangulation [26].

If all the parameters of the stereo system are exactly known, a depth map of absolute distances can be computed. If not, only relative depths can be obtained. However, even in the former case, practical issues limit the range of depths which can be measured, which is generally between 0 to around 30 meters [26].

Given the lack of data sets of stereoscopic images suitable for scene classification, in this paper depth maps were estimated using a method based on single images. We point out that such methods provide only relative depth maps, and exhibit two main drawbacks with respect to methods based on stereo image pairs: a higher computational cost which makes them unsuitable in most real application scenarios, and a lower accuracy. Nevertheless, they are suitable to the purposes of this work, namely to investigate whether depth map information can be useful to discriminate between indoor and outdoor images. Among the different approaches proposed so far to estimate the depth map from a single image [26], we used the one in [15]. It is based on extracting several small image segments, and in inferring their 3D-orientation and 3D-location using Markov Random Fields. This method provides a depth map with a resolution of 55×305 pixels, independent of the original image resolution. It is able to estimate the depth map in 1-2 minutes, depending on the size of the input image.

4 Feature Extraction from Depth Maps

In this section we propose three possible feature sets which can be extracted from depth maps, to discriminate between indoor and outdoor images. We also discuss how such features can be combined with the ones proposed in other works (see Sect. 2), to improve their discriminant capability.

As explained in Sect.3, we consider relative depth maps obtained by the method of [15]. An example is given in Fig. 1.

To define the feature sets, we first analysed the average histogram of relative depth values computed over all images of the two data sets described in Sect. 5,

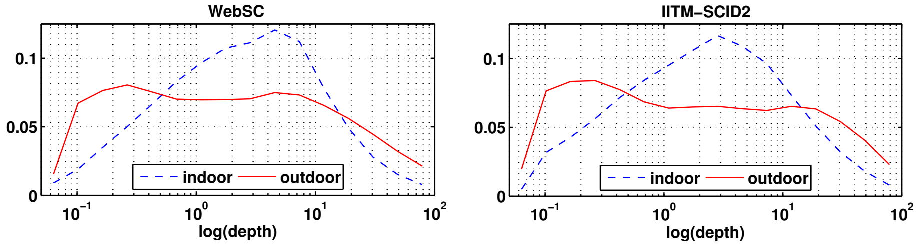


Fig. 2. Average distribution of depth values of images of the WebSC (left) and IITM-SCID2 (right) data set

separately for indoor and outdoor images. These distributions are reported in Fig. 2. It can be seen that indoor and outdoor images exhibit a clearly different behaviour, especially at lower depth values (which are emphasised by the log-scale). Depths of an indoor scene are likely to lie at medium values (see Fig. 1, left), while in outdoor scenes they are distributed more uniformly (see Fig. 1, right). Interestingly, these distributions are very similar over the two image data sets, despite the corresponding images strongly differ in terms of image quality, size, and acquisition device.

The above analysis suggests that a simple set of features potentially exhibiting a discriminant capability between indoor and outdoor images is the histogram of the logarithm of relative depth values of a given image. We denoted this feature set as $3D_{\mathrm{H}}$; its size equals the number of histogram bins.

A drawback of the histogram computed over the whole image is that it does not retain any information about the spatial distribution of depth values. To address this problem, a possible solution is to subdivide an image into $N \times N$ sub-blocks, and to compute the average logarithm of depth values of each sub-block. This feature set is denoted as $3D_{\mathrm{B}}$, and its size is equal to N^2.

Like any other 2D signal, the depth map can be represented in terms of frequency and phase values. Intuitively, outdoor scenes should exhibit an higher contribution at lower frequencies than indoor scenes. Indeed, lower frequencies are likely to correspond to larger homogeneous areas like sky, sea or sandy beach. Based on this intuition, we define a third feature set made up of the average DCT coefficients, computed by a $K \times K$ window sliding over the image. The size of the resulting feature set, named $3D_{\mathrm{D}}$, is K^2.

5 Experimental Evaluation

In this section we experimentally assess the discriminant capability of the feature sets proposed in Sect. 4. We first compare their discriminant capability with the reference methods mentioned in Sect. 2 [13,17,21,27]. We then investigate whether the performance of each reference method can be improved, by using each of the proposed feature sets as additional information source.

5.1 Experimental Setup

Experiments were carried on two benchmark data sets of indoor and outdoor images. Depth maps were generated using the method in [15]. The first data set is IITM-SCID2.[1] It was used in [10,21], and is made up of 907 images (442 indoor, 465 outdoor), subdivided into 393 training and 514 test images. We removed one test image which was too small to be processed by the depth map estimation method. The second data set, denoted as "Web Scene Collection" (WebSC), is made up of 1917 images (955 indoor, 962 outdoor) collected by the authors from the Web and manually labelled. Both data sets, together with the corresponding depth maps, are available at http://prag.diee.unica.it/public/datasets.

The characteristics of the images in the two data sets are rather different. The WebSC corpus contains mainly good-quality, high resolution images. IITM-SCID2 is instead mostly made up of low-quality, low-resolution images, which are often out of focus, and exhibit chromatic aberrations. IITM-SCID2 is thus more challenging than WebSC.

The methods in [17,21] were implemented as follows. They both adopt a two-stage classification scheme. In [17] SVM classifiers with RBF kernel were used at both stages, while in [21] neural networks were used. However, we adopted SVMs with a RBF kernel also for the latter method, as their performance was better than the one of neural networks. To combine the feature sets proposed in this work to the reference methods, we simply added another SVM classifier with a RBF kernel to the first stage, with our features as input.

We used a SVM with a RBF kernel also for the *Gist* [13] and *Centrist* [27] feature sets. To add our feature sets, a two stage classifier similar to the previous ones was built, using again SVMs with a RBF kernel.

Our $3D_{\mathrm{H}}$ feature set was computed as a 16-bin histogram, while for the $3D_{\mathrm{B}}$ feature set images were subdivided into 4×4 sub-blocks. For the $3D_{\mathrm{D}}$ feature set, we chose a sliding window of 8×8 pixel.

We trained the second-stage classifiers of all the methods using the scores provided by first-stage classifiers on training images, obtained through a 5-fold cross-validation. The C parameter of the SVM learning algorithm and the σ parameter of the RBF kernel $K(x_i, x_j) = exp(-\|x_i - x_j\|/(2\sigma))$ were estimated by a 3-fold cross validation on training data. SVMs were implemented using the LibSVM software library [5].

Classification performance was measured as overall and per-class accuracy. Concerning the IITM-SCID2 corpus, we kept the original subdivision into training and testing sets, to allow a direct comparison with the results reported in [10,21]. For WebSC we adopted the 5×2 cross-validation approach of [1], to evaluate also statistical significance. In this case the classifier parameters were estimated separately on each training fold.

As a preliminary evaluation of the degree of complementarity of our feature sets with the ones of the reference methods, we analysed the joint distribution of the score values provided by the corresponding classifiers. Fig 3 shows the joint score distribution for $3D_{\mathrm{B}}$ and *Centrist*, on WebSC. The scores exhibit a high

[1] http://www.cse.iitm.ac.in/~sdas/vplab/SCID/

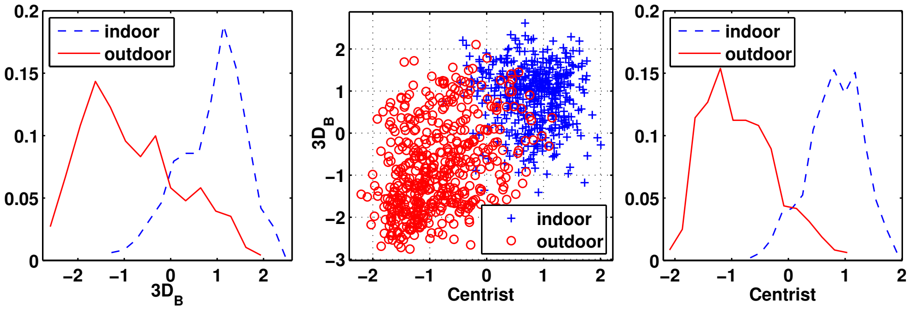


Fig. 3. Joint distributions of the classifier scores obtained by the *Centrist* and $3D_\mathrm{B}$ feature sets on WebSC. Each point in the middle plot corresponds to a single image.

degree of complementarity, which suggests that combining depth information with pixel-based information could improve classification performance. A similar behaviour was observed using all the other feature sets.

5.2 Results and Discussion

Table 1 shows the classification accuracy attained by our proposed features sets, by the four reference feature sets, and by all the possible combinations of one of our feature sets with one of the reference methods.

Our feature sets attained an overall classification accuracy between about 70% and 80%. This supports our intuition that depth map provides useful information to discriminate between indoor and outdoor images. Among the proposed features, $3D_\mathrm{B}$ attained the best performance on both data sets. This suggests that taking into account also the spatial depth distribution (as in the $3D_\mathrm{B}$ features) is beneficial for the considered classification task.

Classifiers based on our features attained however a lower performance than each of the reference methods. Nevertheless, it is worth pointing out that their performance turned out to be comparable to the average accuracy attained by the *individual* feature sets used in [17,21]. In particular, the colour-based features of [17] exhibited an overall classification accuracy of 0.77 and 0.83 respectively on the IITM-SCID2 and WebSC data set, while texture-based features attained an overall accuracy on the same data sets respectively equal to 0.75 and 0.77. The accuracy of the seven feature sets of [21] were between 0.68 and 0.80 for the WebSC data set, and between 0.52 and 0.78 for the IITM-SCID2 data set.

Despite the overall classification accuracy attained by the four reference methods was higher than the one attained by classifiers based on our feature sets, Table 1 shows that the performance of the reference methods was almost always improved when the corresponding classifiers were combined with the ones based on our features. The only exception can be observed for the combination of the *Centrist* and $3D_\mathrm{D}$ features, on the IITM-SCID2 data set.

Table 1. Classification accuracy attained by our proposed features sets (top three rows), by the reference methods (top row of each subsequent group of rows), and by all the possible combinations of each reference method with each of our feature sets (remaining rows). For the WebSC data set, the average accuracy and the standard deviation over the 5×2 cross-val. procedure is reported; $*$ and $**$ denote, respectively, results significant with 90% and 95% confidence, with respect to the f-test.

Method	IITM-SCID2			WebSC			
	Indoor	Outdoor	Total	Indoor	Outdoor	Total	
$3D_{\mathrm{B}}$	0.803	0.742	0.772	0.857 ± 0.011	0.750 ± 0.018	0.803 ± 0.010	
$3D_{\mathrm{H}}$	0.635	0.788	0.713	0.835 ± 0.043	0.745 ± 0.018	0.790 ± 0.018	
$3D_{\mathrm{D}}$	0.695	0.773	0.735	0.813 ± 0.022	0.758 ± 0.019	0.785 ± 0.015	
Centrist [27]	0.960	0.875	0.916	0.932 ± 0.016	0.907 ± 0.007	0.920 ± 0.007	
Centrsti + $3D_{\mathrm{B}}$	0.944	0.909	0.926	0.960 ± 0.009	0.917 ± 0.012	0.938 ± 0.005	**
Centrist + $3D_{\mathrm{H}}$	0.940	0.909	0.924	0.956 ± 0.007	0.910 ± 0.010	0.933 ± 0.006	**
Centrist + $3D_{\mathrm{D}}$	0.920	0.902	0.910	0.954 ± 0.011	0.890 ± 0.021	0.922 ± 0.006	
Tao [21]	0.896	0.811	0.852	0.936 ± 0.007	0.906 ± 0.017	0.921 ± 0.009	
Tao + $3D_{\mathrm{B}}$	0.908	0.864	0.885	0.941 ± 0.011	0.912 ± 0.008	0.927 ± 0.003	*
Tao + $3D_{\mathrm{H}}$	0.863	0.879	0.871	0.945 ± 0.008	0.907 ± 0.010	0.926 ± 0.006	
Tao + $3D_{\mathrm{D}}$	0.871	0.886	0.879	0.938 ± 0.012	0.910 ± 0.010	0.924 ± 0.005	
Gist [13]	0.847	0.852	0.850	0.924 ± 0.011	0.876 ± 0.025	0.900 ± 0.014	
Gist + $3D_{\mathrm{B}}$	0.884	0.860	0.871	0.938 ± 0.011	0.891 ± 0.015	0.914 ± 0.009	*
Gist + $3D_{\mathrm{H}}$	0.851	0.883	0.867	0.941 ± 0.013	0.885 ± 0.022	0.913 ± 0.014	*
Gist + $3D_{\mathrm{D}}$	0.847	0.883	0.865	0.944 ± 0.009	0.862 ± 0.021	0.903 ± 0.010	
Serrano [17]	0.871	0.837	0.854	0.871 ± 0.014	0.866 ± 0.010	0.868 ± 0.005	
Serrano + $3D_{\mathrm{B}}$	0.876	0.883	0.879	0.912 ± 0.013	0.870 ± 0.012	0.891 ± 0.009	**
Serrano + $3D_{\mathrm{H}}$	0.855	0.883	0.869	0.908 ± 0.018	0.872 ± 0.012	0.890 ± 0.012	**
Serrano + $3D_{\mathrm{D}}$	0.827	0.894	0.862	0.896 ± 0.010	0.879 ± 0.017	0.887 ± 0.007	**

To assess the statistical significance of the observed accuracy improvements on the WebSC data set, we performed the f-test of [1]. We did not apply this test on the IITM-SCID2 data set, as a single run of the experiments was carried out on the predefined subdivision into a training and a testing set. Table 1 shows that the accuracy improvements were found to be statistically significant at the 90% or 95% confidence level, in eight out of twelve cases. In particular, the improvements attained by using the $3D_{\mathrm{B}}$ feature set turned out to be always statistically significant with a confidence level of at least 90%. These results provide evidence that, although image depth information may exhibit a lower discriminant capability than other information sources for the indoor/outdoor image classification task, it also provides *complementary* information, as suggested by the results reported at the end of Sect. 5.1. This can allow to improve the discriminant capability of other information sources, by combining them with image depth information.

6 Conclusions and Future Work

In this work we investigated the usefulness of image depth map as an information source in the task of scene classification, and in particular to discriminate between indoor and outdoor images. Our interest is motivated by the rapid diffusion of stereoscopic image acquisition devices which is expected in the next years. We provided evidences that relative depth maps embed in their ”structure” useful information for discriminating between indoor and outdoor images. Moreover, we showed that such information exhibits a complementariness to other information sources used by state-of-the-art methods, and that the discriminant capability of the latter can be improved by combining their feature sets with features extracted from depth maps.

We point out that our experiments were carried out by estimating relative depth maps from single images, due to the current lack of data set of stereo images. Since estimated depth maps are less accurate than the ones which can be obtained from stereo images, the latter can be expected to provide an even higher discriminant capability than the one observed in our experiments.

An interesting follow-up of this work is to devise more informative features based on depth maps, as the ones considered in this work are only a preliminary attempt. To this aim, the information about spatial distribution of depth values could be further investigated, since it exhibited the highest discriminant capability among the considered features. It is also interesting to investigate the usefulness of depth information for other, more complex scene classification tasks. Finally, it will be clearly useful to construct a data set of stereo images, representative of a real application scenario.

Acknowledgements. This work was partly supported by a grant from Regione Autonoma della Sardegna awarded to Ignazio Pillai, PO Sardegna FSE 2007-2013, L.R.7/2007 “Promotion of the scientific research and technological innovation in Sardinia”.

References

1. Alpaydin, E.: Combined 5 x 2 cv F test for comparing supervised classification learning algorithms. Neural Computation 11(8), 1885–1892 (1999)
2. Battiato, S., Farinella, G.M., Gallo, G., Ravi, D.: Exploiting Textons Distributions on Spatial Hierarchy for Scene Classification. EURASIP Journal on Image and Video Processing (2010)
3. Bianco, S., Ciocca, G., Cusano, C., Schettini, R.: Improving color constancy using indoor-outdoor image classification. IEEE Trans. on Image Processing 17(12), 2381–2392 (2008)
4. Boutell, M., Luo, J.: Beyond pixels: Exploiting camera metadata for photo classification. Pattern Recognition 38(6), 935–946 (2005)
5. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines (2001), http://www.csie.ntu.edu.tw/~cjlin/libsvm/

6. Collier, J., Ramirez-Serrano, A.: Environment classification for indoor/outdoor robotic mapping. In: Canadian Conference on Computer and Robot Vision, CRV 2009, pp. 276–283 (May 2009)
7. Deng, D., Zhang, J.: Combining multiple precision-boosted classifiers for indoor-outdoor scene classification. In: Int. Conf. on Information Technology and Applications, vol. 2, pp. 720–725 (2005)
8. Ehinger, K.A., Torralba, A., Oliva, A.: A taxonomy of visual scenes: Typicality ratings and hierarchical classification. Journal of Vision 10(7), 1237 (2010)
9. Fei-Fei, L., Iyer, A., Koch, C., Perona, P.: What do we perceive in a glance of a real-world scene? Journal of Vision 7(1) (2007)
10. Gupta, L., Pathangay, V., Patra, A., Dyana, A., Das, S.: Indoor versus outdoor scene classification using probabilistic neural network. EURASIP Journ. on Adv. in Signal Processing, Special Issue on Image Perception 2007(1), 123–123 (2007)
11. Heitz, G., Gould, S., Saxena, A., Koller, D.: Cascaded classification models: Combining models for holistic scene understanding. In: NIPS (2008)
12. Lee, B.N., Chen, W.Y., Chang, E.Y.: A scalable service for photo annotation, sharing, and search. In: Proc. of the 14th Annual ACM Int. Conf. on Multimedia, pp. 699–702 (2006)
13. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. Jour. of Computer Vision 42, 145–175 (2001)
14. Payne, A., Singh, S.: Indoor vs. outdoor scene classification in digital photographs. Pattern Recognition 38(10), 1533–1545 (2005)
15. Saxena, A., Sun, M., Ng, A.: Make3d: Depth perception from a single still image. In: Proc. of The AAAI Conf. on Artificial Intelligence, pp. 1571–1576 (2008)
16. Schettini, R., Brambilla, C., Cusano, C., Ciocca, G.: Automatic classification of digital photographs based on decision forests. International Journal of Pattern Recognition and Artificial Intelligence 18(5), 819–845 (2004)
17. Serrano, N., Savakis, A., Luo, J.: A computationally efficient approach to indoor/outdoor scene classification. In: ICPR, vol. 4, pp. 146–149 (2002)
18. Serrano, N., Savakis, A.E., Luo, J.: Improved scene classification using efficient low-level features and semantic cues. Pattern Recognition 37(9), 1773–1784 (2004)
19. Sinha, P., Jain, R.: Classification and annotation of digital photos using optical context data. In: Proc. of The 2008 Int. Conf. on Content-Based Image and Video Retrieval, New York, NY, USA (2008)
20. Szummer, M., Picard, R.W.: Indoor-outdoor image classification. In: Proc. of IEEE Int. Workshop on Content-Based Access of Image and Video Database, pp. 42–51 (1998)
21. Tao, L., Kim, Y.H., Kim, Y.T.: An efficient neural network based indoor-outdoor scene classification algorithm. In: Int. Conf. on Consumer Electronics (ICCE). Digest of Technical Papers, pp. 317–318 (2010)
22. Torralba, A., Oliva, A.: Semantic organization of scenes using discriminant structural templates. In: Int. Conf. on Computer Vision, pp. 1253–1258 (1999)
23. Torralba, A., Oliva, A.: Depth estimation from image structure. IEEE Trans. on Pattern Analysis and Machine Intelligence 24 (2002)
24. Tversky, B., Hemenway, K.: Categories of environmental scenes. Cognitive Psychology 15(1), 121–149 (1983)
25. Vailaya, A., Figueiredo, M., Jain, A., Zhang, H.J.: Image classification for content-based indexing. IEEE Trans. on Image Processing 10(1), 117–130 (2001)
26. Wei, Q.Q.: Converting 2d to 3d: A survey (2005)
27. Wu, J., Rehg, J.M.: Centrist: A visual descriptor for scene categorization. IEEE Trans. on Pattern Analysis and Machine Intelligence 99 (2010)

A Multiple Component Matching Framework for Person Re-identification

Riccardo Satta, Giorgio Fumera, Fabio Roli, Marco Cristani,
and Vittorio Murino

Dept. of Electrical and Electronic Engineering,
University of Cagliari Piazza d'Armi, 09123 Cagliari, Italy
{riccardo.satta,fumera,roli}@diee.unica.it
Istituto Italiano di Tecnologia (IIT)
Via Morego 30, 16163 Genova, Italy
{marco.cristani,vittorio.murino}@iit.it

Abstract. Person re-identification consists in recognizing an individual that has already been observed over a network of cameras. It is a novel and challenging research topic in computer vision, for which no reference framework exists yet. Despite this, previous works share similar representations of human body based on part decomposition and the implicit concept of multiple instances. Building on these similarities, we propose a Multiple Component Matching (MCM) framework for the person re-identification problem, which is inspired by Multiple Component Learning, a framework recently proposed for object detection [3]. We show that previous techniques for person re-identification can be considered particular implementations of our MCM framework. We then present a novel person re-identification technique as a direct, simple implementation of our framework, focused in particular on robustness to varying lighting conditions, and show that it can attain state of the art performances.

Keywords: person re-identification, multiple instance learning, framework.

1 Introduction

In video surveillance, person re-identification is the task of recognizing an individual that has already been observed over a network of cameras. It is a novel research topic with many challenging issues, like low resolution frames, different and time-varying light conditions, and partial occlusions.

So far, no theoretical framework for the person re-identification problem exists yet. Previous works are based on different, apparently unrelated approaches, and are focused on devising effective features and appearance models. Despite this, most works turn out to adopt a similar *part-based* body representation, and/or use implicitly the concept of *multiple instances* [15] by considering image patches, regions, or points of interest.

It is worth to note that the above commonalities among previous works bear a resemblance with a framework recently proposed for object detection, Multiple

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 140–149, 2011.

Component Learning (MCL) [3]. In fact, MCL adopts a part-based multiple-instance object representation: an object is considered as a sequence of parts, and each part is represented as a set of instances. A Multiple Instance Learning (MIL) approach is then used to recognise each individual part, using positive and negative examples of instances of that part. Despite this analogy, the MCL framework can not be directly applied to person re-identification, which is a *recognition* rather than a *detection* task. Moreover, only one or a few positive examples (the template images of a given person) are usually available, while no negative examples are generally considered. This makes person re-identification a task more suited to a matching approach rather than a recognition approach based on learning of human body models. In fact, most previous works formulated person re-identification as a task where template images of individuals are collected and then matched against probe images by some similarity measure.

Based on the above premises, in this work we propose a framework for person re-identification named Multiple Component Matching (MCM). We embed in MCM the part-based multiple-instance body representation approach underlying both MCL and previous works on person re-identification, with the aim to provide a framework which can be used as a reference to develop new methods, and possibly to improve current ones. We show at first that techniques proposed in previous works can be seen as particular implementations of MCM. Then, we present a novel person re-identification method, as a more direct and simple implementation of MCM, focused in particular on robustness to variations of lighting conditions, which attains state of the art performances.

We overview MCL and previous works on person re-identification in Sect. 2. The MCM framework is presented in Sect. 3, where its relationships with previous works are also discussed. Our implementation of MCM is presented in Sect. 4, and is experimentally evaluated in Sect. 5. In Sect. 6 future research directions are discussed.

2 Background and Previous Works

Here we present first the MCL framework, and then overview previous works on person re-identification.

2.1 Multiple Component Learning

Multiple instance learning (MIL) is a general learning paradigm for problems in which samples are made up by a *bag* (set) of labeled *instances*, and only the label of the whole bag is known. The task is to build a classifier which learns to label bags using the feature vectors of their instances [15]. MIL has been applied to several computer vision problems, including scene classification [11], image retrieval [16], and object detection [13].

MCL is an extension of MIL, tailored to object detection [3]. In MCL an object is represented as a *sequence of sets of components* (bags of instances in MIL terminology). The rationale behind is the independent detection of different object components, to gain robustness to partial occlusions. Moreover, the subdivision in components is not predefined, but obtained as part of the learning process. To this aim, the image is first subdivided into a predefined set of randomly chosen regions, and a classifier is trained on each region to detect the corresponding component. Since the corresponding object component may appear in different positions inside a region, each region is randomly subdivided into a set of possibly overlapping subregions (*patches*), only some of which may contain the component. A MIL classifier is then used to detect the corresponding component, threating each patch as a single instance. An ensemble of MIL classifiers, one for each region, is trained via a boosting algorithm, so that the most discriminative regions get a higher weight. Learning is thus done both at the *set* level (to detect individual object components) and at the *sequence* level (to combine the information on object components coming from different regions). Note that, according to the MIL paradigm, in the training set only the image regions are labeled, while the patches inside each region are not.

2.2 Previous Works on Person Re-identification

Person re-identification consists in associating an individual from a probe set to the corresponding template in a gallery set. Depending on the number of available frames per individual, the following scenarios can be defined [5]: *Single vs Single* (SvsS), if only one frame per individual is available both in probe and in gallery sets; *Multiple vs Single* (MvsS), if multiple template frames per individual are available in the gallery set; *Multiple vs Multiple* (MvsM), if multiple frames per individual are available both in the probe and gallery sets. The most challenging scenario is SvsS.

All the above scenarios have been considered in [5]. Here, human body is subdivided with respect to its symmetry properties: anti-symmetry separates head, torso and legs, while symmetry is exploited to divide left and right parts. The descriptor is made up of three local features: colour histograms for torso and legs, weighted with respect to the distance from the symmetry axis; *maximally stable colour regions* (MSCR) and *recurrent high-structured patches* (RHSP), both extracted from torso and legs separately. To obtain MSCR and RHSP, several patches are sampled at random, mainly near symmetry axes; then, clustering algorithms are used to find the most significant ones. The matching distance is a combination of the distances computed on the individual features. In MvsS and MvsM scenarios, templates are accumulated into a single descriptor.

In [2], an human body parts detector is used to find in the body of each individual fifteen non-overlapping square cells, that have proven to be "stable regions" of the silhouette. For each cell a covariance descriptor based on colour gradients is computed. Descriptor generation and matching is performed through a pyramid matching kernel. This method can be applied only to SvsS scenarios.

In [1] two methods were proposed. In the first, Haar-like features are extracted from the whole body, while in the second the body is divided into upper and lower part, each described by the MPEG7 Dominant Colour descriptor. Learning is performed in both methods, respectively to choose the best features and to find the most discriminative appearance model. The training set for each individual consists of different frames as positive examples (MvsS, MvsM scenarios), and of everything which is not the object of interest as negative examples. In the SvsS scenario (one frame per individual), different viewpoints are obtained by sliding a window over the image in different directions.

An approach based on harvesting SIFT-like interest points from different frames of a video sequence is described in [9]. Different frames are used also in [6], where two methods are proposed. The first one is based on interest points, selected on each frame by the Hessian-Affine interest operator. The second one exploits a part subdivision of the human body based on decomposable triangulated graphs and dynamic programming to find the optimal deformation of this model for the different individuals. Each part is then described by features based on colour and shape; the distance between a template and a probe is a combination of the distances between pairs of corresponding parts.

In [8] the problem of defining the best descriptor for person re-identification is addressed. Different features are extracted, and their weights are computed by a boosting algorithm. Features are computed from randomly taken strips.

In [12] person re-identification is considered as a relative ranking problem, exploiting a discriminative subspace built by means of an Ensemble RankSVM. Colour and texture-based features are extracted from six fixed horizontal regions.

Despite the methods summarised above exhibit many differences, it can be noted that all of them are based on some part-based body representation, and/or exploit more or less implicitly the concept of multiple instances. This provides the foundation for the proposed framework, which is depicted in the next section.

3 A Framework for Person Re-identification

In this section we describe the proposed Multiple Component Matching (MCM) framework for person re-identification. As mentioned previously, MCM is inspired by MCL; in fact, we found that the concepts behind most previous work are similar to the ones underlying MCL, namely part subdivision and multiple component representation.

Like in MCL, an object is represented as an ordered sequence of sets. In turn, each set is made up by several components. Differently from object detection problems addressed by MCL, we view person re-identification as an object recognition problem where a matching approach is used without any learning phase: while in the training samples of MCL a set is composed by both negative and positive components (the first contain the object part of interest, while the latter do not), in MCM only positive ones are available, namely only those corresponding to body parts of the template person. Formally, let $\mathcal{T} = \{\mathbf{T}_1, \ldots, \mathbf{T}_N\}$ be the *gallery set* of templates, each corresponding to an individual. Each template $\mathbf{T}_i$ is represented by an ordered sequence of a predefined number of M sets, corresponding to the M parts into which an image is subdivided:

$$\mathbf{T}_i = \{T_{i,1}, \ldots, T_{i,M}\} \tag{1}$$

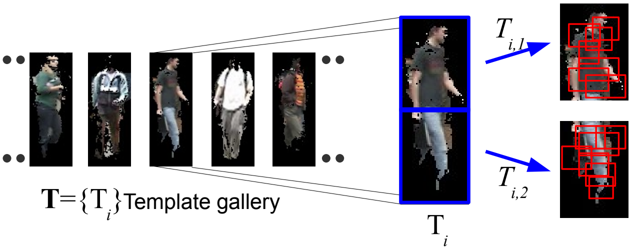


Fig. 1. Representation of an individual according to MCM. Each template $\mathbf{T}_i$ of the gallery $\mathcal{T}$ is represented by an ordered sequence of M parts $T_{i,j}$ (here $M = 2$, corresponding to upper and lower body parts). Each part is made up of a set of components (here, rectangular patches, in red). A feature vector $t^k_{i,j}$ describes each component.

Following a multiple-instance representation, every part $T_{i,j}$ is represented by a set of an arbitrary number $n_{i,j}$ of elements (instances in MIL, components in MCL) (see Fig. 1), and is described by the corresponding feature vectors $\mathbf{t}^k_{i,j}$:

$$T_{i,j} = \{\mathbf{t}^1_{i,j}, \ldots, \mathbf{t}^{n_{i,j}}_{i,j}\}, \mathbf{t}^k_{i,j} \in \mathbb{X}, \tag{2}$$

where $\mathbb{X}$ denotes the feature space (assumed the same for all sets, for the sake of simplicity, and without losing generality). Given a probe $\mathbf{Q}$, which is represented as a sequence of parts as described above, the task of MCM is to find the most similar template $\mathbf{T}^* \in \mathcal{T}$, with respect to a similarity measure $D(\cdot, \cdot)$:

$$\mathbf{T}^* = \arg\min_{\mathbf{T}_i} D(\mathbf{T}_i, \mathbf{Q}). \tag{3}$$

We consider a similarity measure D between sequences defined as a combination of similarity measures $d(\cdot, \cdot)$ between sets:

$$D(\mathbf{T}_i, \mathbf{Q}) = f\big(d(T_{i,1}, Q_1), \ldots, d(T_{i,M}, Q_M)\big). \tag{4}$$

Similarly to MCL, where learning is performed both at the sequence level and at the set level, in MCM the two similarity measures D and d are defined at sequence and at the set level. The similarity measure between sequences D can be any combination of the set distances, like a weighted average in which the coefficients reflect the relevance of the corresponding regions. The following considerations can be made on the choice of a proper similarity measure d between sets. In MCL, this level corresponds to build a MIL classifier. In the MIL paradigm, only a subset of the instances belonging to a set may be responsible of the label of the whole set. Analogously, in MCM a template set can be considered to match the corresponding probe set, if at least a few pairs of instances of the two sets are "similar": the object components which may be in common to the two parts represented by the sets can be placed, in fact, everywhere inside the parts, and therefore be "captured" by any of the instances.

Accordingly, the similarity measure between sets can be defined as the minimum of the similarity measures between all pairs of their instances. To minimize the sensitivity to outliers, more complex measures can be defined, for example considering the first k best matches instead of only one. Moreover, the similarity measure can take into account relationships among instances in a set, i.e. the relative spatial disposition of the corresponding components.

Previous Works and MCM

Most of the previous works on person re-identification described in Sect. 2.2 can be framed into MCM.

A common way to face the kinematics of the human body is to adopt a division in parts [1,2,5,6,12], consistently with the part subdivision of MCM. In MCM, every part is made up of several components. In [5], two of the three features proposed are based on such subdivision: both MSCR and RHSP represent multiple patches of the considered body part. When parts are not represented by components, as in the third feature of [5], and in [1,2,6,12], this can still be considered as a special case of MCM, where every part is composed by only one component. On the contrary, while in [1,6,8,9] no part-based body representation is used, a multiple-instance representation is nevertheless adopted, in the form of interest points or patches. These methods can be seen as particular implementations of MCM as well, where only one body part is considered.

In the definition of MCM, we assumed that every part is represented in the same feature space. This is however not a strict requirement. If different feature vector representations are used, a part can be represented by several feature sets. Accordingly, also the methods in [5,8], where several feature vector representations were used, can be seen as MCM implementations.

4 A Method for Person Re-identification Based on MCM

We propose here a novel person re-identification method as a possible example of a direct implementation of the MCM framework. In particular, our method exploits MCM to attain robustness to changing illumination conditions.

We assume that the "blob" of the person has been already extracted by some detector, and thus only pixels belonging to the mask of the blob are considered. To divide the body into parts, we exploit the anti-symmetry axes as proposed in [5] (see Fig. 2) to locate torso and legs. The head is discarded, since it does not carry enough information due to its relatively small size.

Every body part is represented as a set of a fixed number P of rectangular patches of random area in the range $[12.5\%, 25\%]$ of the region area. Every patch is described by a pair (HSV, y_{pos}), where HSV is the concatenation of H, S and V histograms of the patch (24, 12 and 4 bins respectively) and y_{pos} is the relative vertical position of the center of the patch, with respect to the height of the region.

As stated in Sect. 3, a challenging issue in person re-identification is how to face lighting changes. MCM suggests a way: the multiple instance representation, in fact, can naturally be exploited here, adding instances corresponding to different lighting conditions. However, such instances can not easily been obtained

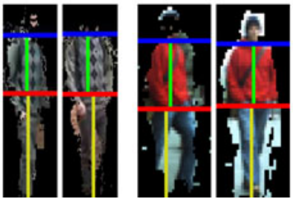


Fig. 2. Body partition: symmetry (vertical) and anti-simmetry (horizontal) axes

from real data (multiple frames, corresponding to as much as possible different illuminations, including lighting gradients and shadows, should be acquired). So, we *simulate* them by constructing artificial patches from real template ones.

Light variations usually result in a change of both brightness and contrast of the image (see for example Fig. 3-a). Brightness variations can be obtained by adding or subtracting a fixed value to the RGB components of the pixels of the image. Instead, changing contrast means increasing or decreasing the differences between pixel values. A standard method to obtain this is the following: denoting as $[0, C]$ the original range of each colour channel (usually $C = 255$), every R, G, and B pixel value is translated to $[-C/2, C/2]$, multiplied by a fixed coefficient, and then re-normalised to $[0, C]$. A coefficient greater than 1 results in a higher contrast, while a lower contrast is obtained by choosing values smaller than 1.

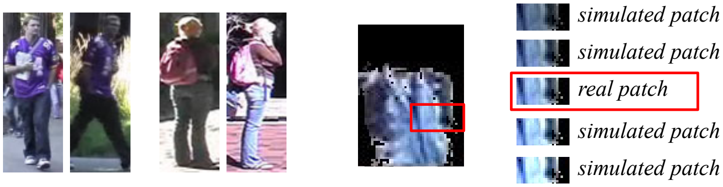
To change both brightness and contrast, we propose a modification of the above method, which does not translate values to $[-C/2, C/2]$ first, but simply multiplies each pixel value of each channel by a coefficient K. Intuitively, this increases (or decreases) the differences between pixel values as well. However, while in the standard method values lower than $C/2$ are reduced, and those higher than $C/2$ are increased, in our variant all the values are increased (or decreased), thus obtaining also a change of brightness. Our algorithm multiplies pixel values by a series of coefficients $[k_1, \ldots, k_S]$ to generate S simulated patches from each real one (see the example in Fig. 3-b). To choose proper k_i values, we start from an initial vector $K = [k_1, \ldots, k_S]$, then decrease its values until applying the greatest k_i to the original image does not saturate the image too much. More precisely, we check that the mean value of R, G and B multiplied by the greatest value of K is not higher than a threshold, which we set to 240.

As explained in Sect. 3, to implement MCM two similarity measures D and d have to be defined. D was defined as the average of the distances between sets. Concerning d, in set theory a common distance function between sets is the *Hausdorff Distance* [4], defined as the maximum of the minimum distances between each element of one set and each element of the other. Comparing two sets $X = \{x_i\}$ and $Y = \{y_i\}$, we have

$$d_H(X, Y) = \max(h(X, Y), h(Y, X)) \tag{5}$$

where

$$h(X, Y) = \max_{x \in X} \min_{y \in Y} (\|x - y\|) \tag{6}$$



(a) (b)

Fig. 3. (a) Two examples of different images for the same individual: note the difference both in contrast and brightness. (b) Examples of four artificial patches simulating changing illumination (right), corresponding to the patch highlighted on the left.

Such a distance measure is sensitive to outlying elements. To avoid this issue, we adopted the *k-th Hausdorff Distance* proposed by Wang and Zucker [14], which takes the k-th ranked distance rather than the maximum: in Eq. 5, in place of $h(X, Y)$ we have then

$$h_k(X, Y) = \underset{x \in X}{kth} \min_{y \in Y}(\|x - y\|) \tag{7}$$

Finally, to compute the norm $\|x-y\|$ in Eq. 7, a distance metric must be defined for the pairs (HSV, y_{pos}) that describe each patch. Denoting with $b(HSV_1, HSV_2)$ the Bhattacharyya distance between histograms, we defined the metric as

$$(HSV_1, y_{pos,1}) - (HSV_2, y_{pos,2}) = b(HSV_1, HSV_2) \cdot (1 + \beta)|y_{pos,1} - y_{pos,2}|$$

where β controls the relevance of the difference in spatial position of the patches.

The above method is applicable only to SvsS scenarios. To extend it to MvsS and MvsM scenarios, one possible approach is to accumulate the patches over different frames in the same sequence of sets.

5 Experimental Results

The performance of the proposed MCM implementation was assessed on the VIPeR benchmark dataset [7], a challenging corpus composed by two non overlapping views of 632 different pedestrians, which show varying changing conditions and pose variations. The best performing method so far is SDALF [5].

We evaluated performance in terms of cumulative matching characteristics (CMC) curve, which represents the probability of finding the correct match over the first n ranks. As in [5], we obtained blob masks by the STEL generative model [10]. In our method we set $\beta = 0.6$, used $P = 80$ real patches, and adopted $K = [1.4, 1.2, 1.0, 0.8, 0.6]$ as the initial vector coefficients for simulation. The value of k for the k-th Hausdorff Distance measure was set to 10.

We employed the same experimental setup of [5], to obtain comparable results. Ten random subsets of 316 pedestrians were drawn from the original dataset. The gallery set is composed by the first image of each person; the probe set,

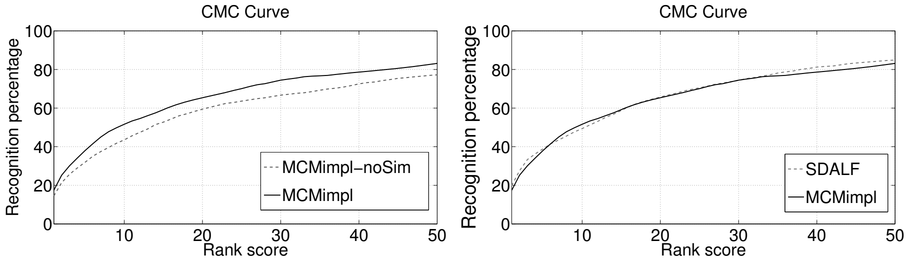


Fig. 4. Performances on the ViPER dataset. (left) Our MCM implementation (*MCMimpl*) with and without simulation. (right) *MCMimpl* compared with *SDALF*.

by the second one. Images of the probe set are compared to the images of the gallery set to find the best match. We used the same images as in [5].

In Fig. 4(left), we reported the CMC curve attained by our method with and without simulation. Simulating varying lighting conditions allows to attain a much better performance, at the expense of a higher overall computation time which is due to template creation and to the higher number of patches compared.

In Fig. 4(right) we compare the performance of our method, including the simulation, with SDALF. As shown, the proposed implementation of MCM attains a performance which is close to the reference. What make $MCMimpl$ highly preferable is the computational cost: SDALF involves time-consuming operations like clustering, transformations, cross-correlation, etc., while $MCMimpl$ performs only simple and fast operations.

The average computation time of our proposed method on a 2.4 GHz CPU is reported in Tab. 1. The method was implemented in C++, without any particular optimization or parallelization. As a qualitative comparison, the implementation of SDALF made available by the authors of SDALF, written partly in C++ and partly in MATLAB, requires over 13 seconds to build one descriptor, and performs a match in around 60 ms, on the same 2.4 GHz CPU. Note that these computational times can not be directly compared to the ones reported for $MCMimpl$, due to the partial MATLAB implementation. However, reimplementing MATLAB code in a more performing language usually results in a speed-up of no more than 10-20 times, pretty far from the difference of 2 orders of magnitude between the descriptor creation times.

Table 1. Average computation time per frame or, for matching, per pair of frames

	Template desc. creation	Probe desc. creation	Matching
MCM_{impl}	$93.7ms$	$6.8ms$	$28.6ms$
$MCM_{impl,nosim}$	$6.8ms$	$6.8ms$	$6.7ms$

6 Conclusions and Future Work

We proposed a framework for person re-identification which embeds common ideas underlying most of the previous works, and is inspired by the MCL framework for object detection. We also developed a simple MCM implementation including a method to make it robust to changing illumination conditions, which has a low computational cost and attains a performance close to the state-of-the-art method on a benchmark data set.

Two main directions for further research can be foreseen. First, simulation can be implemented in MCM to attain robustness also to pose variations. Second, it is interesting to investigate whether MCM can be extended to a learning approach. Exploiting its relationships with MCL, this can enable MCM to adopt a MIL approach to learn the appearance of each body part, as an alternative to the matching approach considered here, taking advance of the large available literature on that learning framework.

References

1. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Person re-identification using haar-based and dcd-based signature. In: AVSS, pp. 1–8 (2010)
2. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Person re-identification using spatial covariance regions of human body parts. In: AVSS (2010)
3. Dollár, P., Babenko, B., Belongie, S., Perona, P., Tu, Z.: Multiple component learning for object detection. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 211–224. Springer, Heidelberg (2008)
4. Edgar, G.A.: Measure, Topology, and Fractal Geometry (1990)
5. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR (2010)
6. Gheissari, N., Sebastian, T.B., Hartley, R.: Person reidentification using spatiotemporal appearance. In: CVPR (2006)
7. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: PETS (2007)
8. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008)
9. Hamdoun, O., Moutarde, F., Stanciulescu, B., Steux, B.: Interest points harvesting in video sequences for efficient person identification. In: VS (2008)
10. Jojic, N., Perina, A., Cristani, M., Murino, V., Frey, B.: Stel component analysis: Modeling spatial correlations in image class structure. In: CVPR, pp. 2044–2051 (2009)
11. Maron, O., Ratan, A.L.: Multiple-instance learning for natural scene classification. In: ICML (1998)
12. Prosser, B., Zheng, W., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: BMVA, pp. 21.1–21.11 (2010)
13. Viola, P.A., Platt, J.C., Zhang, C.: Multiple instance boosting for object detection. In: NIPS (2005)
14. Wang, J., Zucker, J.D.: Solving the multiple-instance problem: A lazy learning approach. In: ICML (2000)
15. Yang, J.: Review of multi-instance learning and its applications. Tech. Rep. (2005)
16. Zhang, D., Wang, F., Shi, Z., Zhang, C.: Interactive localized content based image retrieval with multiple-instance active learning. Patt. Rec. 43(2), 478–484 (2010)

Improving Retake Detection by Adding Motion Feature

Hiep Van Hoang[1], Duy-Dinh Le[2], Shin'ichi Satoh[2], and Quang Hong Nguyen[1]

[1] Hanoi University of Science and Technology, No 1 Dai Co Viet,
Hanoi City, Vietnam
[2] National Institute of Informatics, 2-1-2 Hitotsubashi,
Chiyoda-ku, Tokyo, Japan

Abstract. Retake detection is useful for many applications of video summarization. It is a challenging task since different takes of the same scene are usually of different lengths; or have been recorded under different environment conditions. A general approach to solve this problem is to decompose the input video sequence into sub-sequences and then group these sub-sequences into clusters. By combining with temporal information, the clustering result is used to find take and scene boundaries. One of the most difficult steps in this approach is to cluster sub-sequences. Most of previous approaches only use one keyframe for representing one sub-sequence and extract features such as color and texture from this keyframe for clustering. We propose another approach to improve the performance by combining the motion feature extracted from each sub-sequence and the features extracted from each representing keyframe. Experiments evaluated on the standard benchmark dataset of TRECVID BBC Rushes 2007 show the effectiveness of the proposed method.

1 Introduction

In film and video production, people usually have to deal with large amount of un-edited or rushes video content since one scene in rushes of video might have been recorded many times for different reasons. Each recording of the scene involves different takes, and only the best take is selected to make the final edited content while the others are removed. The takes that are removed are called redundant takes or retakes of the selected take, and finding the takes that are belong to one scene to group them into one cluster is the problem of retake detection. One motivation of retake detection is it can help editors by automatically finding the best takes to be used. Another motivation is in problem of video summarization, where editors have to make a shorter version of original video.

There are several challenges when solving the problem of retake detection:

– Different takes of the same scene usually have different lengths of time. For example, some shots are disrupted by miscues, bloopers, or various unexpected mistakes.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 150–157, 2011.

- Even if two takes whose duration are very similar are still slightly different due to differences in environmental conditions or various differences in the actions of actors. If only several keyframes are used for segment representation, it will be sensitive to the acquiring conditions such as changing light, and changing positions of objects or actors.
- The boundary of retakes and scene formed by these takes is usually unknown.

Bailer et al. [1] proposed a distance measure based on Longest Common Subsequence (LCSS) to determine the similarity of segments in the video. These similar segments were then grouped into one cluster if they belong to one scene. Dumont and Merialdo [2] divided the input video into one-second segments and then used Smith-Waterman algorithm to find all similarity segments in the video. Feng Wang and Chong-Wah Ngo [3] tried to find takes directly by using shot boundary detection combined with speech recognition.

Most of these approaches only focus on matching video sequences in some feature space. In this paper, we focus on another aspect that is how to represent video sequences (i.e. feature representation of video sequences) to improve the matching performance. We propose to use motion feature that was not considered in related work to encode spatio-temporal relation between consecutive frames. By adding the motion feature to existing features such as color and texture, the representation has more discriminative power, resulting better matching performance.

2 Framework Overview

We adopt the general framework proposed in [2] for the problem of retake detection in video (c.f Figure 1). First, the input video is divided in segments by using shot boundary detector [1,3] or sampling every 1-second length [2]. Next, keyframe images are extracted from the segments and then features are extracted from these keyframe images. A clustering method such as k-means or hierarchical clustering is used to group similar segments into clusters. After the clustering step, by using the cluster labels, the input video is represented as a list of labels.

Since one retake might contain several segments, Smith-Waterman algorithm is then applied to find all repeated sub-sequences [2]. This step is considered as a further grouping of segments to form retakes. Finally, scenes that are sets of retakes are formed. We use rand index metric proposed by W. M. Rand [4] to evaluate the results against the groundtruth.

2.1 Feature Extraction

Color Moments (CM)
Color moments have been successfully used in retrieval systems [5] and proved to be efficient and effective in representing color distributions of images [6].

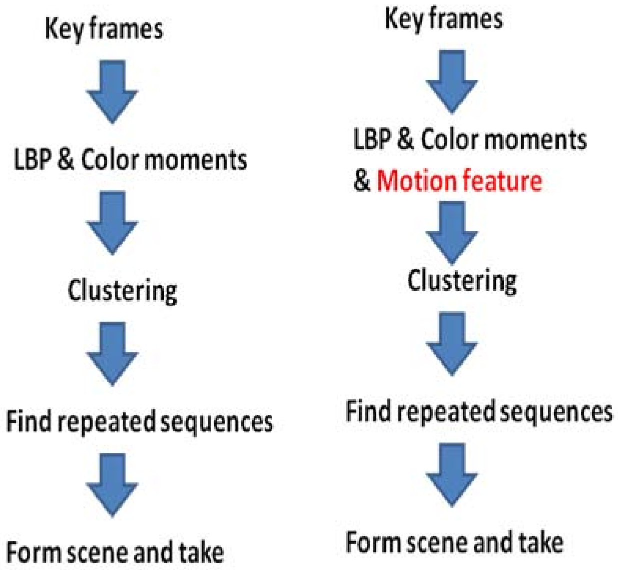


Fig. 1. An overview of a general framework for the problem of retake detection. (left) The framework used in [2]. (right) Our framework improve the performance of retake detection by adding motion feature.

The first order (mean), the second order (variance) and the third order (skewness) color moments are defined as:

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} f_{ij}$$

$$\sigma_i = (\frac{1}{N}\sum_{j=1}^{N}(f_{ij} - \mu_i)^2)^{\frac{1}{2}}$$

$$s_i = (\frac{1}{N}\sum_{j=1}^{N}(f_{ij} - \mu_i)^3)^{\frac{1}{3}}$$

where f_{ij} is the value of the i-th color component of the image pixel j, and N is the number of pixels in the image.

Local Binary Patterns (LBP)

The LBP operator proposed by Ojala et al. [7] is a powerful method for texture description. It is invariant with respect to monotonic grey-scale changes, hence no grey-scale normalization needs to be done prior to applying the LBP operator. This operator labels the pixels of an image by thresholding the neighborhoods of each pixel with the center value and considering the result as a binary number.

Motion Feature

Since optical flow is an approximation of image motion based on local derivatives of sequence images. It specifies how much each pixel moves between adjacent frames in the video. If I(x, y, t) is the center in $(m \times m)$ neighbourhood and moves by δx, δy in time δt to I(x + δx, y + δy, t + δt). Since I(x, y, t) and I(x + δx, y + δy, t + δt) are intensity of the same pixel, we have:

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t) \tag{1}$$

Assuming the movement to be small, we can perform a 1^{st} Taylor series expansion about I(x, y, t) as follows:

$$I(x + \delta x, y + \delta y, t + \delta t) = I(x, y, t) + \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t + H.O.T.$$

Where H.O.T. are Higher Order Terms, which we assume are small and can be ignored. Following the constrain (1) above, we must have:

$$\frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t = 0$$

or

$$\frac{\partial I}{\partial x}\frac{\delta x}{\delta t} + \frac{\partial I}{\partial y}\frac{\delta y}{\delta t} + \frac{\partial I}{\partial t}\frac{\delta t}{\delta t} = 0$$

which results in

$$\frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t}\frac{\delta t}{\delta t} = 0 \tag{2}$$

Where V_x and V_y are the x and y components of the velocity or optical flow of I(x, y, t) and $\frac{\partial I}{\partial x}$, $\frac{\partial I}{\partial y}$, and $\frac{\partial I}{\partial t}$ are the derivatives of the image at (x,y,t) in the corresponding direction. We normally write these partial derivatives as: $I_x = \frac{\partial I}{\partial x}$; $I_y = \frac{\partial I}{\partial y}$, and $I_t = \frac{\partial I}{\partial t}$, finally the equation (2) can be rewritten as

$$(I_x, I_y).(V_x, V_y) = -I_t \tag{3}$$

We cannot compute V_x and V_y directly since (3) is an equation in two unknowns. This is known as the aperture problem of the optical flow algorithms, so we applied Lucas and Kanade method [8] to solve this problem. After the optical flow of two consecutive keyframe images was computed, an orientation histogram of optical flow is built and used as the motion feature.

To discriminate two different keyframe images more accurately, the keyframe images are divided into regions by using a 5 × 5 grid. The features are extracted for each region and concatenated to form the final feature vector. Then, all features including color moments, local binary patterns, and motion features are concatenated and normalized to a zero mean and unit standard deviation.

2.2 Finding All Repeated Sub-Sequences

We use k-means clustering for grouping segments. After that, the input video is presented as a list of strings $S = s_1 s_2 \ldots s_n$ where s_i is the cluster label of the segment represented by keyframe image f_i in the video and n is the number of keyframes extracted from the video. Since the video contains several scenes and each scene is recorded several times (each time is called a take), several sub-sequences of S should be repeated at different positions in the video sequence. Our work is to find these repeated sub-sequences. To do so, we adopt the method described in [2] that uses Smith-Waterman algorithm [9]. This algorithm is based on dynamic programming and is designed to not only compare the global alignment of two sequences of characters (two strings) but also their local alignment. To find all aligned sub-sequence of two sequences, a scoring matrix of $(m+1) \times (n+1)$ is located with each column for each character in the first sequence and each row for each character in the second sequence. Cells of scoring matrices indicate the cost of changing a sub-sequence of the first sequence to a sub-sequence of the second sequence.

Build Scoring Matrix. Given two sequences of characters: $A = a_1 a_2 \ldots a_m$, $B = b_1 b_2 \ldots b_n$. A $(m+1) \times (n+1)$ scoring matrix H is built as:

$$H(i,0) = 0; 0 \leq i \leq m$$

$$H(0,j) = 0; 0 \leq j \leq n$$

$$H(i,j) = max \begin{bmatrix} & 0 & \\ H(i-1,j-1) + w(a_i,b_j), & & Match(Mismatch) \\ & H(i-1,j) + w(a_i,-), & Deletion \\ & H(i,j-1) + w(-,b_j), & Insertion \end{bmatrix}$$

$$(1 \leq i \leq m; 1 \leq j \leq n)$$

where $w(a_i, b_j)$ is the match/mismatch score, if $a_i = b_j$ then $w(a_i, b_j) = w(match)$ else $w(a_i, b_j) = w(mismatch)$. H(i,j) is the score of similarity between two sequences that end at a_i, b_j respectively. The m is the length of sequence A and the n is the length of sequence B.

Figure 2 below shows an example of the scoring matrix when two sequences $S_1 = ACACACTA$ and $S_2 = AGCACACA$ are compared. In this example, w(match) = +2; w(mismatch) = w(a,-) = w(-,b) = -1.

Trace Back to Find Optimum Local Alignment. To obtain the optimum local alignment of two sequences, we start at the highest value in scoring matrix (i,j) and then trace it back from this position to one of three previous positions $(i-1, j-1)$, $(i-1; j)$, and $(i, j-1)$ that depend on which position has a maximum value. This trace back is repeated until we meet a matrix cell with a zero value. In the example above, the optimum local alignment starts at position (8, 8) and the trace back sequence is (7,7), (7,6), (6,5), (5,4), (4,3), (3,2), (2,1),

$$H = \left(\begin{array}{cccccccccc}
 & - & A & C & A & C & A & C & T & A \\
- & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
A & 0 & 2 & 1 & 2 & 1 & 2 & 1 & 0 & 2 \\
G & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\
C & 0 & 0 & 3 & 2 & 3 & 2 & 3 & 2 & 1 \\
A & 0 & 2 & 2 & 5 & 4 & 5 & 4 & 3 & 4 \\
C & 0 & 1 & 4 & 4 & 7 & 6 & 7 & 6 & 5 \\
A & 0 & 2 & 3 & 6 & 6 & 9 & 8 & 7 & 8 \\
C & 0 & 1 & 4 & 5 & 8 & 8 & 11 & 10 & 9 \\
A & 0 & 2 & 3 & 6 & 7 & 10 & 10 & 10 & 12
\end{array}\right)$$

Fig. 2. An example of a scoring matrix

(1,1), and (0,0); The final result of two sub-sequences A-CACACTA of sequence 1 and AGCACAC-A of sequence 2 are aligned together with a confident score of 12.

Smith-Waterman Algorithm. We use Smith-Waterman algorithm to find all repeated sub-sequences as follows:

– Input: Given video $S = s_1 s_2 \ldots s_n$
– Build scoring matrix H of two sequences: S and itself
– Loop :
  - Find the position (i, j) in which H(i, j) is maximum
  - If H(i, j) < threshold break;
    ∗ Trace back to find optimum aligned sub-sequence and store it (two sub-sequences are aligned together if they are belong to trace back trajectory)
    ∗ Update scoring matrix to find next aligned sub-sequence
– Output: a rank list of aligned sub-sequences (after this, it is called a list of take candidates)

2.3 Forming Takes and Scenes

We have a list of take candidates in this step. Our work now is grouping all takes of the same scene into one group and each group will be a scene. We do the following steps to form a scene and a take.

– Remove all take candidates whose length is too short (< 4 seconds)
– Sort take candidates in time order
– Two take candidates $p_1 = [start_1, end_1]$ and $p_2 = [start_2, end_2]$ are merged into one take $p = [min(start_1\ , start_2),\ max(end_1, end_2)]$ if

- p_1 and p_2 overlap by time: $p_1 \cap p_2 \neq \phi$, and
- $Length(p_1 \cap p_2) > 0.5 * min(length(p_1, p_2))$

– Two take candidates are grouped into one if they are aligned together.

Figure 3 visualizes our algorithm to form scenes and takes. The first line is the ground truth of the video. There are two scenes in this example that have been visualized in blue and green. The next lines are take candidates that are sorted in time order; two takes that are aligned together have the same color. The solid colored rectangles represent take candidates; the open dotted rectangles represent takes after candidates have been merged.

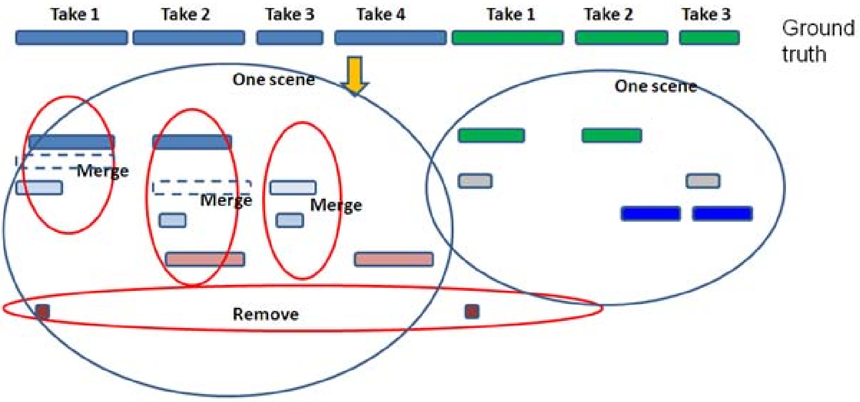


Fig. 3. Illustration on how to form takes and scenes

3 Experiments

We have tested our proposed method on five videos selected randomly from TRECVID BBC Rushes 2007. The groundtruth is provided by NIST that was used for evaluation of the summarization task in TRECVID benchmark 2007. Rand index [4] is used to evaluate the performance of the two methods without and with using motion feature. Since the performance of the methods depends on choosing the best k in k-means clustering, we tried different values $k = 180, 190, 200, 210, 220, 230, 250$ and only the best performance is reported in the comparison table.

As shown in table 1, adding motion feature boosts the detection performance both in scenes and takes. We also found that motion feature does not help improve performance in all situations. The MRS157475 video is an example. The reason is this video was recorded outdoors in high wind conditions. In this case, the motion feature is not useful since it has a lot of noises and current motion feature could not handle.

Table 1. Experimental result on 5 videos of TRECVID 2007. Performance of each method is evaluated using rand index (RI) score. The higher RI score, the better performance.

Videos	**Features**			
	CM&LBP		CM&LBP&Motion	
	Scene RI	Take RI	Scene RI	Take RI
MRS044500	0.58	0.60	0.73	0.73
MS216210	0.69	0.71	0.66	0.70
MRS157475	0.73	0.81	0.70	0.73
MRS025913	0.60	0.63	0.67	0.69
MRS144760	0.62	0.64	0.71	0.72
Average	0.64	0.68	**0.69**	**0.71**

4 Conclusions

Retake detection is a important but challenging task for applications of video summarization. Existing work [2,1] only focuses on how to group sub-sequences into takes and scenes. In order to improve further the performance, we propose to add motion feature in current features such as color and texture used in these existing frameworks. Experimental results on BBC Rushes dataset of TRECVID 2007 show effectiveness of the proposed method.

References

1. Bailer, W., Lee, F., Thallinger, G.: A distance measure for repeated takes of one scene. The Visual Computer: International Journal of Computer Graphics 25, 53–68 (2008)
2. Dumont, E., Merialdo, B.: Rushes video summarization and evaluation. Multimedia Tools and Applications 48, 51–68 (2010)
3. Wang, F., Ngo, C.W.: Rushes video summarization by object and event understanding. In: TVS 2007 Proceedings of the International Workshop on TRECVID Video Summarization, pp. 25–29 (2007)
4. Rand, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66, 846–850 (1971)
5. Flickner, M., Sawhney, H.S., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., Yanker, P.: Query by image and video content: The qbic system. IEEE Computer 28, 23–32 (1995)
6. Stricker, M.A., Orengo, M.: Similarity of color images. In: Proc. of SPIE, Storage and Retrieval for Image and Video Databases III, vol. 2420, pp. 381–392 (1995)
7. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. on Pattern Analysis and Machine Intelligence 24, 971–987 (2002)
8. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of Imaging Understanding Workshop, pp. 121–130 (1981)
9. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. Journal of Molecular Biology 147, 195–197 (1981)

RDVideo: A New Lossless Video Codec on GPU

Piercarlo Dondi[1], Luca Lombardi[1], and Luigi Cinque[2]

[1]Department of Computer Engineering and Systems Science, University of Pavia,
Via Ferrata 1, 27100 Pavia, Italy
[2]Department of Computer Science, University of Roma ”La Sapienza”,
Via Salaria 113, Roma, Italy
{piercarlo.dondi,luca.lombardi}@unipv.it,
cinque@di.uniroma1.it

Abstract. In this paper we present RDVideo, a new lossless video codec based on Relative Distance Algorithm. The peculiar characteristics of our method make it particularly suitable for compressing screen video recordings. In this field its compression ratio is similar and often greater to that of the other algorithms of the same category. Additionally its high degree of parallelism allowed us to develop an efficient implementation on graphic hardware, using the Nvidia CUDA architecture, able to significantly reduce the computational time.

Keywords: Lossless Video Coding, Screen Recording, GPGPU.

1 Introduction

Recording screen activity is an important operation that can be useful in different fields, from remote descktop applications to usability tests. In the first case is essential to reduce the amount of trasferred data, it is not strictly necessary a high quality video compression: a lossy compression and a small set of informations (e.g. the list of pressed keys or of started programs) are generally sufficient[1]. On the contrary, in the second case, it is generally useful to have a complete recording of the screen in order to study the user interaction.

There are many screen recording system on market [2] that implement different features, but all of them use traditional codec for compress the output.

In this paper we introduce a new method for lossless video coding, called RDVideo, based on Relative Distance Algorithm [3], whose peculiar characteristics make it particularly suitable for compressing screen video recording. In this field its compression ratio is similar and often greater to that of the other algorithms of the same category, like Lagarith [4], HuffYUV [5], FFV1 [6] or Alpary [7].

Additionally, its high degree of parallelism can be very useful for increasing its performances, in particular with a GPU implementation. The execution on the Nvidia CUDA architecture gives interesting results: the comparison between CPU and GPU showed the significant speed-up in elaboration time supplied by graphic hardware, both in compression and in decompression.

The paper is organized as follows: section 2 presents the description of the codec; its parallel implementation on CUDA is described in section 3; section

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 158–167, 2011.

4 shows the experimental results, an analysis of compression ratio and a comparison of the performances between the sequential realization on CPU and the parallel on GPU; finally our conclusions in section 5.

2 Relative Distance Video Codec (RDVideo)

2.1 Relative Distance Algorithm (RDA)

The RDA is a lossless method for image compression that uses local operators to minimize the number of bits needed to memorize the data. The base idea of the algorithm is that, in a small area, also the color differences between pixel are generally small [8]. This subsection provides only a general description of the algorithm, with the aim of give the information needed for understanding the extension for the video. A more complete description can be found in a previous work [3]. Every channel of an image is compressed in the same way independently from the others, then we can describe the procedure for a single channel image without loss of generality.

At the beginning the image is divided in blocks of 8x8 pixels. In every block it looks for the maximum (max_p) and the minimum (min_p) pixel value; the found min_p is subtracted to every pixel p_i of the block (1).

$$p'_i = p_i - min_p \tag{1}$$

Each p'_i represents the distance between the element i from min_p – to memorize p'_i is necessary a number of bits less than or equals of that for p_i. We defined the *Relative Distance* (RD) of a block as the minimal number of bits required to represent all the p'_i of that block (2).

$$RD = \lceil log_2 \left(max_p - min_p + 1 \right) \rceil \tag{2}$$

The compression is obtained coding the values p'_i with a number of bits equals to RD; the compression ratio is good only when RD is minor than 7 (RD is always included between [0, 8]), otherwise the data are saved uncompressed.

After this first analysis, every block is further split in four 4x4 sub-blocks and the RDA is applied again to each of these ones for checking if the subdivision can increase the compression ratio. Finally the block is saved with the better configuration.

Different headers identify the type of compression used (one 8x8 block, four 4x4 sub-blocks or no-compression). Every header, except in the no-compression case, contains the minimum (saved with 8 bits) and the RD (saved with 3 bits). The number of bits for a compressed block 8x8 is calculated by (3). Instead the dimension of a subdivided block is the sum of the size of the four 4x4 sub-blocks, each one compressed with different RD (4).

$$NBits_{8x8} = HBits + RD * 64 \tag{3}$$

$$NBits_{4x4} = \sum_{i=1}^{4} HBits_i + (RD_i * 16) \tag{4}$$

$$CR = UBS/NBits \tag{5}$$

The compression ratio (CR) is calculated by (5), where uncompressed block size (UBS) is 512 bits (64 pixels x 8 bits). The sequential implementation requires a time linear with the dimension of the image.

2.2 RDVideo Codec

The RDA gives the best performances when the RD is 0 or 1 (that means large homogeneous color areas, like sea or sky). So, even if generally the medium compression ratio of the algorithm is between 1.5 and 3, with a very uniform image it can grow to 10 or more (the theoretical limit is 42.73 with a monochromatic image). Clearly this situation is not very frequent, but in a video sequence consecutive frames are quite similar, in particular analyzing a screen video recording. Consider some standard tasks – e.g. a user that reads a text or copies a file from a folder to another – in all of them there are small changes between a frame and the next thus the variation of the scene depends only by the user's speed. Making the difference between two consecutive frames we can create a new more uniform image, called difference frame (D-Frame), on which we can successfully apply the Relative Distance Algorithm.

RDA is designed to work only on positive values, so before the compression, a D-Frame must be rescaled (6) adding to all pixels an offset (OF), in the two's-complement arithmetic. The result is saved on 8 bits. This operation can be inverted without loss of information using equation (7). The scale factor must have a value between [0, 256], our experiments show that the better results are achieved when it is between [90, 144].

$$DFrame_i = Frame_i - Frame_{i-1} + OF \tag{6}$$

$$Frame_i = Frame_{i-1} + DFrame_i + OF \tag{7}$$

When there is a change of scene a D-Frame does not give significant improvements. The frames that are very different from previous ones are considered intra-frames (I-Frame) and are compressed directly without other elaborations. Obviously the first frame is always an I-Frame.

The following pseudo code summarizes the main steps of RDVideo codec:

```
save(width, height, framerate);
compressedFrame = compress(frame[0]);
save(compressedFrame);
for (i = 1 to NumFrames)
{
     D-Frame = calcDifference(frame[i], frame[i-1]);
     D-FrameDim = evalCompressionRatio(D-Frame);
     I-FrameDim = evalCompressionRatio(frame[i]);
     if(D-FrameDim < I-FrameDim)
          compressedFrame = compress(D-Frame);
```

```
    else
         compressedFrame = compress(frame[i]);

    save(compressedFrame);
}
```

3 CUDA Parallel Implementation

The modern GPUs offer very high computation capabilities, that can be useful not only for 3D graphic. At the end of 2006, with chipset G80, Nvidia introduced CUDA (Computer Unified Device Architecture), a parallel computing architecture that allows to exploit the potentiality of GPU also in general purpose applications and provides a new parallel programming model and a proper instructions set [9].

3.1 CUDA Specifications

A CUDA GPU is a many core device composed by a set of multithreaded Streaming Multiprocessors (SMs), each of them contains eight Scalar Processor cores (CUDA Cores) for single precision floating-point operations, one unit for double precision, two special function units for transcendental functions and a multithreaded instruction unit. The multiprocessor employs a new architecture called SIMT (single-instruction, multiple-thread), capable to create, manage, and execute hundreds of concurrent threads in hardware with zero scheduling overhead. The GPU acts as a coprocessor for the CPU: the data on which it works must be transmitted from the CPU RAM (the host memory) to the GPU DRAM (the device memory), because the GPU cannot read or write directly on the hard disk. This is the real bottleneck of the system, a good CUDA program should minimize the data transmission.

The GPU has also other levels of memory: each multiprocessor supplies a set of local 32-bit registers and a very fast parallel data cache (or shared memory), a read only constant cache and a texture cache that are shared among all the processors. An appropriate use of all the memory levels can significantly improve the total performances of the program.

The function that is executed on the device by many different CUDA threads in parallel is called ”kernel”. There are two types of kernel: the global ones that are called from the host only, and the device kernel that are called from the device only. A kernel can be executed by multiple equally-shaped thread blocks, so that the total number of threads is equal to the number of threads per block multiplied for the number of blocks. These multiple blocks are organized into a one-dimensional or two-dimensional grid of thread blocks. Thread blocks are executed independently, this allows to schedule them in any order across any number of cores. For this reason the code has a high degree of scalability [9].

Until the last year a GPU was able to execute only one kernel at time, this limitation is overcome by the most recent model of graphic cards that implement the new CUDA FERMI architecture [10].

3.2 RDVideo on CUDA

A CUDA implementation of RDA for a single image was discussed in a previous work [3]. The main difference with the implementation for the video concerns the amount of data to transfer from host to device, in this case very high. So the program has been redesigned, according to the characteristic of the graphic hardware, for minimizing the data transfer and to take advantage from all the memory levels of the CUDA architecture.

Compression. All the compression phases are distributed on three global and one device kernels. Only the final operation, the saving on hard disk, is entrusted to CPU. We start the description analyzing the simplest case, the compression of an I-Frame.

The GPU analyzes each block of the frame, for each of them it chooses the best type of compression (one 8x8 block or four 4x4 sub-blocks) and then compresses them as described in 2.2. The values of the current 8x8 block are saved in shared memory, in this way we can reduce the time needed for multiple accesses to the data; the use of the slower device memory is limited only for the initial reading of input data and for the final writing of the results.

In a CUDA program each block is executed in parallel with each other, there is no way to know the order in which they finish, so is very important the synchronization of the reading and the writing phases in order to avoid any kind of conflict. In this case reading phase is not critical – each processor operates on a different part of the image – more complicate is the management of the writing. The dimension of the compressed data is unknow a priori, it must allocate on the device memory enough space for each block in order to exclude multiple writing on the same memory address. For each frame the GPU has to give back to CPU all the data to be saved: the compressed values, the relative distances, the min_p arrays and a global array of flags that mark the type of compression used for each block. For definition, the last one has always a dimension equals to number of blocks of the image; more complex is determine how much space is needed for the other variables.

Taking advantage of the parallel structure of the graphic hardware we can concurrently calculate the compression for the 8x8 block and for the four 4x4 sub-blocks. The main kernel calls four times (in parallel) a device kernel, while continues the computation for the 8x8 block. The device kernel executes on its data the same operations of the global kernel and gives back the compressed dimension of its sub-block. The results of the two kind of compression are both saved in two different shared memory areas, after the evaluation of the compression ratio only the better value is stored in the output array (Fig. 1).

For the elaboration the GPU saves five RD and five min_p per block, so the system need two arrays with a size of 5*(number of blocks of the images). The dimension of all the arrays must be multiply for the number of channels in case of RGB image. For decreasing the device memory allocation, the compressed data overwrite the correspondent input data that are no longer used. This solution resolves also the issue of concurrent writings: each block writes only the same image area that has previously read.

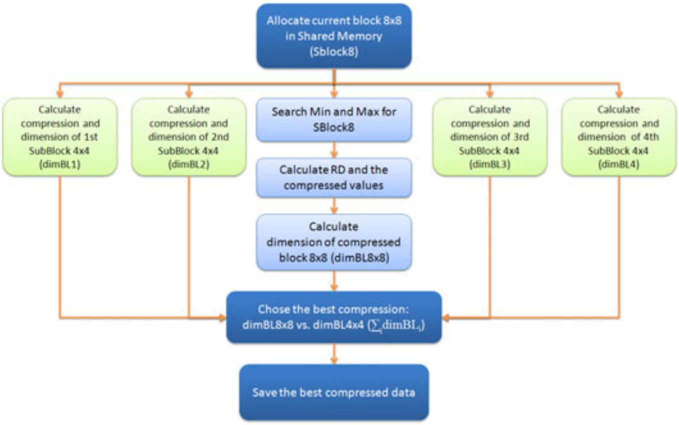

Allocate current block 8x8 in Shared Memory (Sblock8)
Calculate compression and dimension of 1st SubBlock 4x4 (dimBL1)
Calculate compression and dimension of 2nd SubBlock 4x4 (dimBL2)
Search Min and Max for SBlock8
Calculate RD and the compressed values
Calculate dimension of compressed block 8x8 (dimBL8x8)
Calculate compression and dimension of 3rd SubBlock 4x4 (dimBL3)
Calculate compression and dimension of 4th SubBlock 4x4 (dimBL4)
Chose the best compression: dimBL8x8 vs. dimBL4x4 ($\sum_i$dimBL$_i$)
Save the best compressed data

Fig. 1. Flow chart of Compression Kernel for an I-Frame

Summarizing, the steps for compressing an I-Frame are the following:

1. Allocation on device memory of 4 arrays (min_p, RD, flags, image)
2. Copy of the image from host to device
3. Launch of global kernel
4. Copy of all the arrays from device to host

Let us now examine the steps for a generic frame. Firstly a second global kernel creates the D-Frame: the difference between frames is a very simple operation that involves independent data, it fits well on CUDA and it does not required any particular optimization. The obtained D-Frame will be used only by GPU, so it is maintained on device memory and it is not copied on host.

At this point, in the sequential implementation, we check the compression ratio for D-Frame$_i$ and for frame$_i$, in order to calculate the compression only on the one that gives the best CR. This is a good choice for the CPU but not for the GPU, if we first check the dimension of frames and after compress the best one, we must transfer twice the same input data. A better procedure is making the compression during the CR evaluation: a subtraction of independent data is a very fast operation on GPU, so it is more efficient performing a few more operations if it reduces the transfers. Consequently the compression process for a D-Frame follows the nearly steps described in figure 1 for an I-Frame, but the input image is already present on GPU. In the same way works the compression of the frame$_i$.

The trasfer phase requires a specific optimization. The copy on host of both results (compressed D-Frame$_i$ and compressed frame$_i$) is a waste of time; but storing both on the device and transfering only the best one is unsuitable, because with high resolution videos there is an elevated risk to exceed the memory space.

We implement a solution that tries to reach a good balance between memory occupation and the amount of transferred data:

1. Compress D-Frame$_i$.
2. Copy on host all the results (CR and compressed data).
3. Compress frame$_i$.
4. Copy on host only the CR of frame$_i$.
5. If the CR of frame$_i$ is better copy all its data.

Statistically a D-Frame is better than the correspondent I-Frame, so in the most of cases we have only one copy from device to host.

Finally a last optimization: to create a D-Frame the current frame$_i$ and the previous frame$_{i-1}$ are needed, saving the current frame$_i$ in the texture memory it will be already available as frame$_{i-1}$ at next iteration. In addiction accessing the texture memory is faster than reading the device memory. So, making the copy before step 3 of the previous list, also the compression of the frame$_i$ becomes faster. Consequently, as mentioned at the beginning, there is a third global kernel that executes the compression using as input the texture memory rather than the device memory.

Decompression. The decompression is computationally simpler than the compression and also requires less memory transfers. Only the input data must be copied from CPU, the decompressed video can be directly showed on screen by GPU. For this task we need two type of global kernel: one for I-Frames and another for D-Frames. The first simply adds the correspondent minimum to the compressed values of every block. The second makes the same operations and also recreates the original frame$_i$ adding to the just decompressed D-Frame$_i$ the frame$_{i-1}$. To reduce the transfer rate a strategy similar to that used for compression is exploited: the decompressed frame$_i$ is saved in the texture memory to be used again in the next iteration in case of D-Frame. This copy is more efficient than the previous one: for compression the program must send the image from the host memory to the device texture memory, now it makes a copy of the frame$_i$ between two different areas of the GPU memory.

Summarizing, the decompression for a generic frame is composed by four steps:

1. Copy of compressed frame$_i$ from CPU to GPU.
2. Launch of the proper kernel (for D-Frames or for I-Frame).
3. Saving of the decompressed frame$_i$ in texture memory.
4. Visualization on the screen.

4 Experimental Results

The proposed approach is useful for different tasks, but we focused our experiments on screen video recording. In order to test the performances of our algorithm we analyzed different kind of situations, in particular two typical human-computer interactions: a statical use (like reading and writing text with an editor) in which there are very small variations in screen image; and a dynamic use (like opening and moving folders and files on screen) in which the screen changes significantly. The tests are made on an Intel Q9300 CPU with 2.5 GHz and a Nvidia GeForce GTX 260.

The video are recorded at different resolutions (the first five at 1680x1050, the sixth at 1280x1024 and the seventh at 800x600), the duration depends on the type of interaction but it is always in the range of 20-30 sec. Sequence 1, 2, 3 and 7 present a dynamic use, as opening and closing windows and programs or copying files between folders. Test 5 and 6 record a static use, respectively writing with an editor, and playing with a casual game (Solitaire). Finally test 4 considers a particular case, that is partially dynamic and partially static: web navigation with multiple pages visualized.

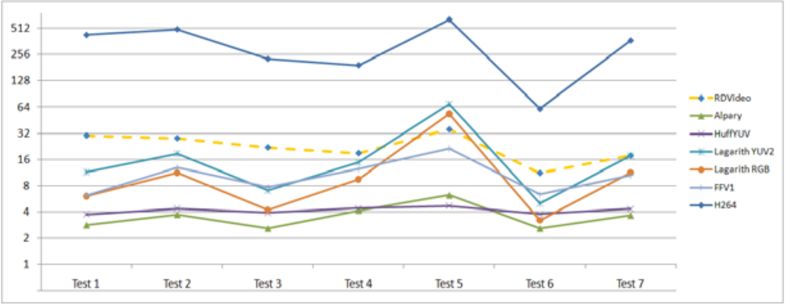


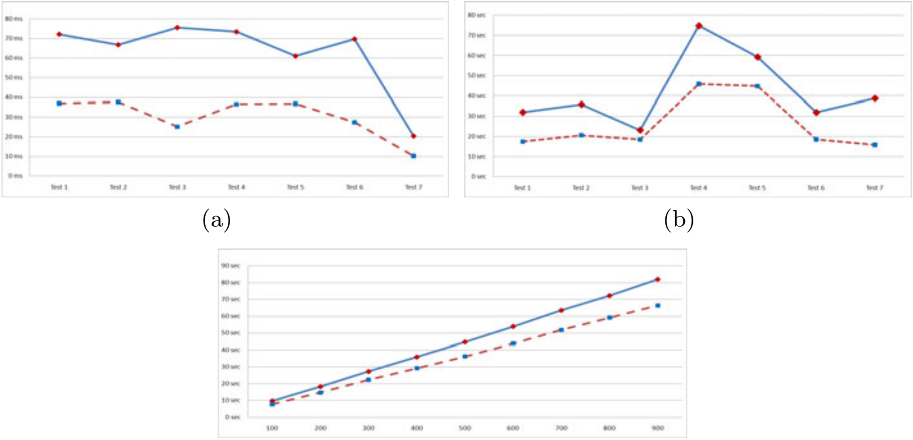
Fig. 2. Compression ratio – comparison between different codecs (RDVideo is the dotted line)

Figure 2 show a comparison of the compression ratio between our method and other standard lossless codecs. In all the videos our method has a good compression ratio, at most only the h264 [11] has a greater compression ratio.

4.1 Execution Times

Compression. For compression we considered the computation time and the total time (computation time plus the time for reading the video input and writing the compressed output). Figure 3(a) shows the medium elaboration time to compress one frame, while figure 3(b) shows the total time to compress the entire sequences. As expected, the outcomes present an evident difference of performances for the computation time in favour of the GPU (Fig. 3(a)).

It must point out that the gap between GPU and CPU is not constant but it is greater in the first five tests where the screen resolution is higher. This is a consequence of the graphic hardware architecture: the GPU has better performances when there are lot of data to analyze, with small frames there is no way to reach the full load of the stream multiprocessors, so a high resolution video fits better than a little one. However, in spite of this improvement, the total time shows a lower enhancement (Fig. 3(b)). This issue, generated by the bottleneck of the data transfer between CPU and GPU, can be overcome considering long recordings. In fact at growing of video lenght the elaboration speed up



(a) (b)

(c)

Fig. 3. Compression: (a) medium elaboration time for one frame (in ms); (b) total time for an entire sequence (in sec); (c) total compression time for the same sequence at growing of number of frames – CPU (continuous line) vs. GPU (dotted line)

supplied by graphic hardware generates a global improvement of performances that reduces, slowly but constantly, the total compression time for the GPU (Fig. 3(c)).

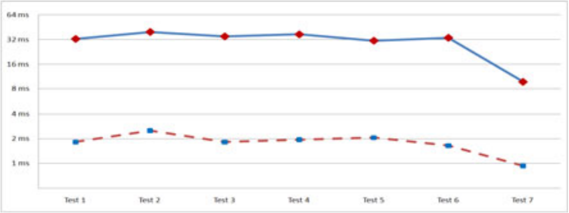


Fig. 4. Medium elaboration time for decompressing one frame: CPU (continuous lines) vs. GPU (dotted lines)

Decompression. For decompression we check only the elaboration time because the total time is function of frame rate, so necessarily they must be similar both for CPU than for GPU. Due to the reduced numbers of operations to execute, the elaboration times are lower than those of the compression, while the difference between CPU and GPU becomes greater (Fig. 4). These outcomes proves that a real-time decompression is always reachable with RDVideo codec, also for high screen resolutions video recordings.

5 Conclusions

In this paper we have presented a new lossless video coding, particularly efficient for the compression of screen video recordings. The experimental results show a compression ratio between 11 and 35 both in static and in dynamic situation, that makes our method a viable alternative in this field to other standard codecs.

The high parallelism of RDVideo codec allowed us to realize an implementation on CUDA architecture that provides a good enhancement of performances for the elaboration time in compression and especially in decompression. The tests have also proved that our method performs better with high screen resolutions and long recordings.

A future update for the codec includes the introduction of a predictor that can be used in addition to D-Frame in order to further improve the compression ratio.

References

1. Shann-Chiuen, W., Shih-Jen, C., Tzao-Lin, L.: Remote Screen Recording and Playback. In: International Conference on Systems and Networks Communications, ICSNC 2006 (2006)
2. Fast, K.: Recording Screen Activity During Usability (2002), http://www.boxesandarrows.com/archives/recording_screen_activity_during_usability_testing.php
3. Bianchi, L., Gatti, R., Lombardi, L., Cinque, L.: Relative Distance Method for Lossless Image Compression on Parallel Architecture. In: Fourth International Conference on Computer Vision Theory and Applications (VISAPP 2009), vol. 1, pp. 20–25 (Febraury 2009)
4. Lagarith Codec, http://lags.leetcode.net/codec.html
5. Huffman, D.A.: A Method for the Construction of Minimum-Redundancy Codes. In: Proceedings of the I.R.E., pp. 1098–1102 (September 1952)
6. FFV1 Codec, http://www1.mplayerhq.hu/~michael/ffv1.html
7. Alparysoft Lossless Video Codec, http://www.free-codecs.com/download/alparysoft_lossless_video_codec.htm
8. Storer, J.A.: Lossless Image Compression Using Generalized LZ1-Type Methods. In: Data Compression Comference (DCC 1996), pp. 290–299. IEEE, Los Alamitos (1996)
9. Kirk, D., Hwu, W.: Programming Massively Parallel Processors: A Hands-on Approach. Elsevier, Amsterdam (2010)
10. Nvidia Corporation, Cuda Programming Guide, http://developer.nvidia.com/object/gpucomputing.html
11. Sullivan, G.J., Topiwala, P., Luthra, A.: The H.264/AVC Advanced Video Coding Standard: Overview and Introduction to the Fidelity Range Extensions. In: SPIE Conference on Applications of Digital Image Processing XXVII (August 2004)

A New Algorithm for Image Segmentation via Watershed Transformation

Maria Frucci and Gabriella Sanniti di Baja

Institute of Cybernetics "E. Caianiello", CNR
Via Campi Flegrei 34, 80078 Pozzuoli (Naples), Italy
{m.frucci,g.sannitidibaja}@cib.na.cnr.it

Abstract. A new segmentation method is presented. The watershed transformation is initially computed starting from all seeds detected as regional minima in the gradient image and a digging cost is associated to each pair of adjacent regions. Digging is performed for each pair of adjacent regions for which the cost is under a threshold, whose value is computed automatically, so originating a reduced set of seeds. Watershed transformation and digging are repeatedly applied, until no more seeds are filtered out. Then, region merging is accomplished, based on the size of adjacent regions.

1 Introduction

Image segmentation is a key process for many image analysis and computer vision tasks. It aims at dividing the image into a number of disjoint regions, ideally corresponding to the regions perceived by a human observer as constituting the scene. While all the pixels in the same region are similar with respect to some property, e.g., color, intensity, or texture, adjacent regions significantly differ from each other with respect to the same property. The result of segmentation is either a collection of regions, constituting a partition of the image, or a collection of contours, each of which delimiting a region of the image.

Segmentation has received much attention in the literature (e.g., refer to [1-9] and to the references quoted therein). Histogram thresholding, edge detection, region growing, fuzzy clustering, probabilistic Bayesian methods, watershed transformation are among the most commonly followed approaches. Selection of the method to be used mostly depends on the specific image domain and application. In this work, we consider watershed-based segmentation.

As far as we know, watershed-based segmentation was introduced in [10]. Then, a number of papers suggesting solutions to the main problems affecting the watershed segmentation, or dealing with different applications of watershed transformation have been published (see, e.g., [11-16]).

Basically, watershed segmentation is based on the identification of suitable *seeds* in the gradient image of the input image, followed by a growing process that originates from the selected seeds. The seeds are generally detected as the sets of pixels with locally minimal value (called *regional minima*). In turn, the growing process associates to each seed the pixels that result to be closer, in terms of a given property, to that seed more than to any other seed.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 168–177, 2011.

A positive feature of watershed segmentation is that the contours delimiting the regions into which the image is divided are mostly placed where human observers perceive them. In fact, the growing process is performed on the gradient image, where the edges are enhanced. A negative feature is that the image may result to be divided into a number of parts that is remarkably either larger (over-segmentation) or smaller (under-segmentation) than the expected number of parts. Over-segmentation is mainly caused by the fact that, by using all the regional minima in the gradient image, a too large number of seeds is obtained. Thus, suitable criteria to reduce the number of seeds are necessary. In turn, under-segmentation occurs when the criteria adopted for filtering the seeds are too selective.

The new method presented in this paper consists of two phases, both aimed at reducing the number of regions into which the image is partitioned to the most significant ones. During the first phase, a reduced set of seeds is computed by means of digging, while during the second phase adjacent regions are merged. The watershed transformation is initially computed starting from all seeds detected as regional minima in the gradient image. The cost for digging a canal to link regional minima of each pair of adjacent regions is computed. Then, digging is actually performed for each pair of adjacent regions for which the cost is under a threshold, whose value is computed automatically. Digging filters out a number of seeds so that a smaller number of regions are obtained when the watershed transformation is applied again. Watershed transformation and digging are repeatedly applied, until the number of seeds, and hence of regions, becomes stable. The second phase of the method is a merging process that is done in one inspection of the image resulting at the end of the first phase, and is based on the size of the regions.

The paper is organized as follows. Section 2 introduces the standard watershed transformation; Section 3 presents the new method; experimental results are discussed in Section 4; finally, concluding remarks are given in Section 5.

2 Watershed Transformation

An easy way to explain how the watershed transformation works is the *landscape* paradigm. A 2D digital image can be interpreted as a terrain elevation map, where the gray level g of a pixel p with coordinates (x, y) is the elevation at position (x, y) of the corresponding 3D landscape. The bottom of each valley of the landscape (called *pit*) is mapped into a connected set of pixels of the 2D image characterized by locally minimal gray level, and the top of each hill (called *peak*) is mapped into a connected set of pixels of the 2D image characterized by locally maximal gray level. Let us suppose that the landscape is immersed into water after its pits have been pierced. Immersion into water causes flooding of the landscape. Valleys with the lowest pits are reached first by the water and the corresponding catchment basins start to be transformed into lakes. When the water in a lake reaches the edge separating the catchment basin of that lake from an adjacent basin, a dam has to be built to prevent the water of the lake to overflow into the adjacent basin. Dam construction continues until the water reaches the highest peaks of the landscape. When this is the case, the top lines of the dams constitute the closed watershed lines, each of which surrounds a catchment basin, and the input image results to be partitioned into a number of catchment basins.

To implement the standard watershed transformation, two different strategies can be followed, known as watershed by topographical distances and watershed by immersion (see [12] for a detailed description of both strategies). In any case, to guarantee that the catchment basins are separated from each other by a leak-proof set of watershed lines, different metrics have to be adopted for basins and lines. We use 4-connectedness for the catchment basins and 8-connectedness for the watershed lines. Connected component labeling completes watershed transformation by assigning an identity label to each of the regions into which the image is partitioned. An example of watershed segmentation is given in Fig. 1. There, image # 42049 of the Berkeley segmentation dataset [17] is to the left and the standard watershed transform is in the middle. Since all 7569 regional minima detected in the gradient image have been used as seeds, the image is partitioned into 7569 regions. The resulting image is clearly over-segmented.

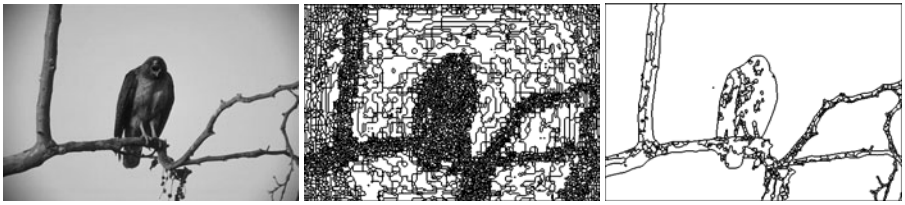


Fig. 1. The gray-level image #42049, left, the standard watershed segmentation in 7569 regions, middle, and segmentation in 645 regions by the algorithm [13], right

To reduce over-segmentation, a careful selection of the regional minima to be used as seeds for region growing is necessary, or reliable criteria for region merging have to be adopted. In [13], an effective but computationally expensive watershed segmentation method has been suggested that significantly reduces over-segmentation. After the watershed transform is computed starting from all seeds, the significance of each obtained region X is evaluated in terms of the interaction of X with every adjacent catchment basin Y. Then, flooding and digging are employed to cause disappearance of the regional minima corresponding to non-significant regions and the watershed transform starting from the reduced set of seeds is computed. Detection of non-significant regions, flooding and digging, and watershed transformation are repeated until all regions in the current watershed transform result to be significant. By applying the watershed segmentation algorithm [13] to the image in Fig. 1 left, the partition into only 645 regions shown in Fig. 1 right is obtained. Unfortunately, the computational cost of this algorithm is rather high, due to the large number of applications of the watershed transformation and, mainly, to the processes aimed at computing region significance.

Flooding and digging, though both aimed at filtering out non-significant seeds, act differently as concerns the resulting watershed transform and should be applied under different circumstances. To understand what happens if the seed associated to a given catchment basin X is removed by either flooding or digging, let us refer to the simple example shown in Fig. 2, where X is shown together with the adjacent catchment basin Y.

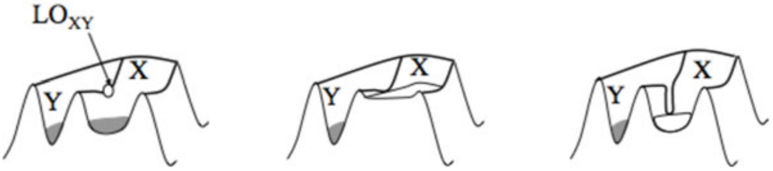


Fig. 2. Local overflow for the catchment basins X and Y. Effect of flooding, middle, and of digging, right.

Let LO_{XY} be the *local overflow pixel*, i.e., the pixel with minimal height along the watershed line separating X from Y and let l_{XY} denote the gray level of LO_{XY}. To filter out the seed associated to X, we should modify the gradient image in such a way that the set of pixels that constituted the seed for X is no longer a regional minimum. To reach the goal via flooding, we must increase to l_{XY} the gray level of all pixels of X with gray level lower than l_{XY}. In turn, to reach the goal by digging, we should dig a canal connecting the regional minima of X and Y. The canal is identified in the gradient image as the minimal length path linking the regional minima of X and Y, and passing through LO_{XY}. The gray level of all the pixels in the path is set to the lower value between those of the regional minima of X and Y.

Clearly, both flooding and digging cause the suppression of the regional minimum associated to X and, hence, X will result as merged to Y when the watershed transformation is applied again. However, digging is generally preferable since it allows to merge X and Y without altering the watershed lines separating X any another adjacent region W with which X should not be merged, even if the local overflow pixel LO_{XW} has gray level l_{XW} smaller that l_{XY}. For this reason, in our new segmentation method we use only digging for seed filtering in the gradient image.

3 The Suggested Method

Our segmentation method consists of two phases, both aimed at reducing the number of regions into which the image has been partitioned by an initial watershed transformation. During the first phase, seed reduction is obtained via digging, while during the second phase adjacent regions are merged if some conditions on their size are verified.

The watershed transformation is initially computed starting from all regional minima in the gradient image. Then, we compute the cost for digging a canal to link the regional minima of each pair of adjacent regions and perform digging for each pair of adjacent regions with cost under a threshold. The value of this threshold is computed automatically. The effect of digging is a reduction in the number of seeds, and hence in the number of regions, that are obtained by applying again the watershed transformation. Watershed transformation and digging are repeatedly applied, as far as seed removal is possible. Then, the second phase is performed. This is a region merging based on the size of the regions and accomplished in one inspection of the image.

3.1 First Phase

As mentioned in the previous section, we prefer to filter out seeds by digging rather than by flooding since digging does not alter the watershed lines separating a region whose minimum has to be suppressed from other regions that should remain in the watershed transform. To perform digging between two adjacent regions X and Y, we need to build a path linking the regional minima of X and Y. The larger is the height of the local overflow pixel LO_{XY} with respect to the regional minima of X and Y, the larger is the cost C_{XY} necessary to build the path and, hence, to perform digging between X and Y passing through LO_{XY}.

If we denote by l_{XY}, min_X and min_Y the gray levels of LO_{XY} and of the regional minima of X and Y respectively, the cost C_{XY} is computed as follows:

$$C_{XY} = (l_{XY}- min_X) + (l_{XY}- min_Y)$$

The contour of a region X consists of a number of parts, each of which is a watershed line separating X from one of the basins adjacent to X. For each region Y adjacent to X, we identify the local overflow pixel LO_{XY} and compute the corresponding cost C_{XY} necessary for digging a canal from X to Y through LO_{XY}.

In general, the watershed lines separating X from the adjacent regions are likely to be characterized by different digging costs. To reduce over-segmentation, we unavoidably need to filter out a number of seeds. Then, it seems safer to perform digging through the local overflow pixel with the smallest associated digging cost, while the remaining watershed lines of X remain unaltered.

However, if the above process is indiscriminately done for all regions in the initial watershed transform, the resulting image would be under-segmented and if digging is iterated the whole image would obviously merge into a unique region.

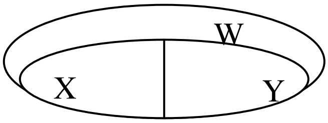


Fig. 3. Three adjacent basins. See text.

We observe that if we want to maintain untouched the watershed lines separating X from other adjacent regions while we dig out a canal linking the regional minimum of X to the regional minimum of Y, some specific conditions should be fulfilled as regards the digging costs along the contour of Y. More in detail, let us refer to Fig. 3 and suppose that we want to dig a canal between X and Y, leaving the remaining part of the contour of X untouched.

Let us consider the costs of digging the canals from X to its adjacent basins Y and W. By taking into account the above issues, we can suppose that for the basin X, C_{XW} is the maximal cost, while C_{XY} is the minimal cost. If we want to preserve the watershed line separating X from W while we merge X and Y via digging, then C_{YW} should be larger than C_{XY}. Otherwise, if C_{YW} is smaller than C_{XY}, besides merging X with Y we would also merge Y to W. As a result, X would be merged to W, while our intention was to keep unaltered the watershed lines separating X from W. Thus, a reasonable condition to prevent digging from X to Y is that C_{XY} is the minimal cost, C_{XW} is the maximal cost, and a basin W adjacent to both X and Y exists such that $C_{YW}<C_{XY}$.

For each maximal cost actually found in the initial watershed transform, we count the number of times that the above situation occurs. In this way, we build the histogram of the distribution of maximal digging costs, where we detect peaks and valleys, i.e., local maxima and local minima. Our aim is to identify a suitable valley whose value can be used as a threshold on digging cost, so as to reduce over-segmentation when the watershed transformation is newly applied. We have experimentally found that, in the average, the best threshold value is the digging cost corresponding to the deepest valley. The depth of a valley is obtained by computing the differences in height between each of the two peaks delimiting any valley and the valley itself, and by taking the largest of the two differences. If the same maximal depth characterizes more than one valley, we use as threshold the digging cost corresponding to the rightmost deepest valley, i.e., we take as threshold value the largest digging cost among those found in all valleys having the maximal depth.

Actually, both very small and rather large maximal cost values are not considered in the histogram to detect the deepest valley. In fact, small digging costs are very frequent in the initial watershed transform, where many noisy partition regions are due to regional minima found in the gradient image in correspondence with almost homogeneous areas. Considering all small digging costs to analyze the histogram would lead to rather high peaks in correspondence with these small costs. This, in turn, would lead to detect a too small threshold value that would result as not adequate to reduce over-segmentation. On the other hand, peak height along the histogram decreases when the digging cost increases, so that the peaks, if any, in the rightmost part of the histogram have rather small height. Considering them would only make histogram analysis more expensive. Moreover, merging between regions for which the digging cost is rather high would mean merging regions that are associated to two regional minima rather different in the gradient image. These regional minima most possibly originate regions in between which a human observer would perceive a separation.

Thus, we build the histogram only for maximal digging costs ranging from a minimum value 10 to a maximal value 40, where both limits have been experimentally found as adequate to provide good segmentation results in the majority of cases.

Once the value of the threshold θ on digging cost has been computed, digging should be done for each pair of adjacent regions X and Y such that $C_{XY}\leq\theta$ (of course, if this condition holds for X and more than just one adjacent region Y, more canals are built, which will produce a larger fusion in the newly computed watershed transform). To this purpose, a linking path passing through LO_{XY} should be built to connect the regional minima of each pair of regions X and Y for which the digging cost does not overcome the threshold. To avoid unwanted digging of X with any other adjacent region W for which a digging cost $C_{XW}>\theta$ is found, the linking path connecting the regional minima of X and Y is not built if the path necessarily passes through pixels that, besides belonging to the watershed line separating X and Y, also belong to the watershed line separating X and W.

The effect of digging is not only that of causing merging of suitable pairs of adjacent regions for which the digging cost is under the threshold, but also that of making possible, at a successive application of the watershed transformation, the construction of linking paths in correspondence with those regions that were characterized by digging cost under the threshold, but for which the paths could not be built. Thus, watershed transformation, computation of the digging costs, and digging are repeated until no more seeds are filtered out. A small number of iterations are generally sufficient to obtain a stable watershed transform. We have experimentally found that in the average the first phase of segmentation requires at most three iterations.

Fig. 4 left shows the result at the end of the first phase of segmentation for the running example. The automatically computed value of the threshold is θ=32, and 666 regions have been found.

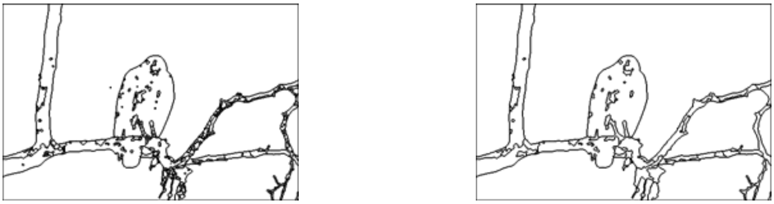


Fig. 4. Result at the end of the first segmentation phase in 666 regions, left, and at the end of the second segmentation phase in 82 regions, right

3.2 Second Phase

The area of each region of the watershed partitioned image is computed by counting the number of pixels it includes. Let A be the arithmetic mean of the areas of all regions. We divide the regions in two classes, respectively including regions with *large* area and regions with *small* area. A region R_i with area A_i is classified as small if it is $A_i/A < \tau$, where the value of the threshold τ has to be fixed depending on problem domain. In this work we have experimentally found that the value τ= 0.33 produces generally good results.

All regions with small area and adjacent to each other are merged into a unique region. For the running example, it results A=231. Due to the selected value for τ, this means that regions with area consisting of less than 77 pixels are regarded as small regions and are accordingly merged.

The segmentation of the running example at the end of the second phase is shown in Fig. 4 right. Only 82 regions characterize the segmented image, which is remarkably less than the 645 regions found by the algorithm [13].

4 Experimental Results

The segmentation algorithm has been tested on a number of images with different resolutions, taken from public databases. The three images #62096, #118035, and #41004 shown in Fig.5, all taken from [17], are used here to show the performance of the segmentation algorithm.

In all cases, the value of the threshold θ has been automatically computed by analyzing the histogram of maximal digging cost distribution in the range from 10 to 40. As for the threshold τ used during the second phase of the process, the value τ=0.33 has been used in all cases. The results obtained at the end of phase 1 and of phase 2 are shown in Fig. 6 top and Fig. 6 bottom, respectively.

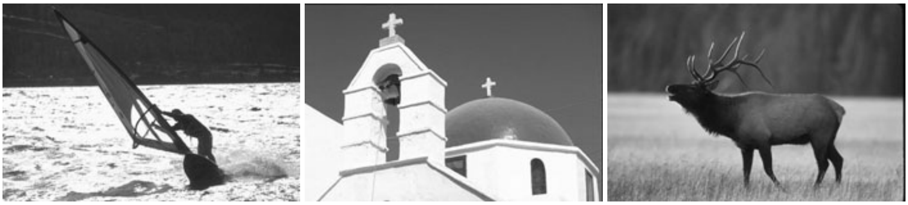


Fig. 5. Test images. From left to right: #62096, #118035 and #41004.

The automatically computed value for the threshold θ is 38, 21, and 20, for images #62096, #118035, and #41004, respectively. Starting from the initial watershed partitions into 14853, 7569 and 9851 regions, at the end of the first phase 2069, 769 and 1966 regions are obtained.

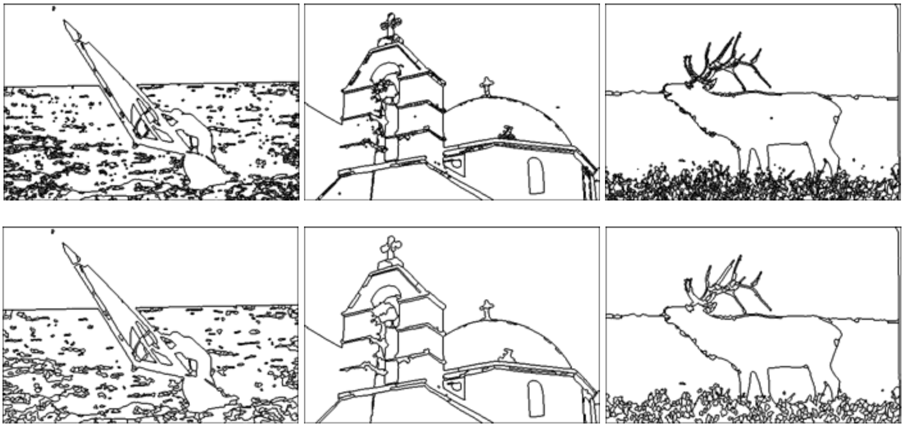


Fig. 6. Results at the end of the first segmentation phase, top, and final segmentations, bottom

In turn, the arithmetic mean of the area of the watershed regions resulting at the end of phase 1 is 74 for #62096, 200 for #118035 and 78 for #41004. As already pointed out, the default threshold value τ= 0.33 has been used for all examples. The final segmentations include 358, 78 and 260 regions, for images #62096, #118035, and #41004, respectively.

We point out that the values 10 and 40 as limits for the analysis of the histogram of the maximal digging costs and the value of the threshold τ=0.33 suggested in this paper have to be interpreted as default values producing, in the average, satisfactory results. Of course, it is not guaranteed that by using the default values the best segmentation is achieved whichever input image is handled. Moreover, we point out that the value of θ automatically computed in correspondence with the deepest valley of the histogram of the maximal digging costs, could generally be slightly increased without producing under-segmented results. As an example refer to Fig. 7, where for image #42049 other two different segmentations, obtained with different values for the segmentation parameters are given. As shown in Fig. 4 right, the value θ=32, automatically computed, produced 666 and 82 regions at the end of the first and of the second segmentation phase. By setting θ=36 (θ=39), 552 and 74 regions (486 and 64 regions) are respectively obtained.

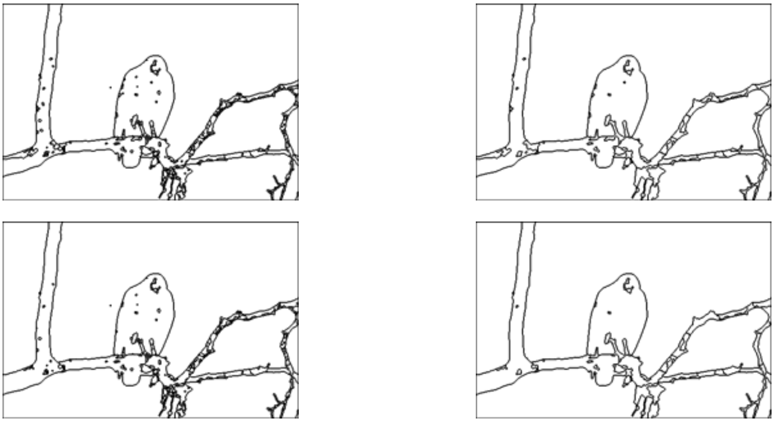


Fig. 7. Different segmentations obtained after the first segmentation phase, left, and after the second phase, right, for θ=36, top, and θ=39, bottom

Analogously, different results are also possible by changing the value of the merging threshold τ used during the second phase of segmentation.

5 Concluding Remarks

We have suggested a segmentation method based on the watershed transformation. To reduce the main drawback of watershed transformation, i.e., the generally too large number of partition regions with respect to the intuitively expected ones, we have introduced the notion of cost of digging. Then, we have established a criterion to compute automatically a threshold on digging cost. For each pair of adjacent basins characterized by a digging cost smaller than the threshold, a path linking the corresponding regional minima has been built so as to merge the two distinct regional minima in one regional minimum. In this way, by applying again the watershed transformation, all regions such that the corresponding regional minima have been connected to each other by means of linking paths result to be merged into a unique region. The whole process consisting of watershed transformation and digging is repeated as far as linking paths can be built to merge adjacent regions. Once the digging cost is above the threshold for all regions, a second phase of segmentation is accomplished to merge adjacent regions based on their relative size. The algorithm has been tested on a large variety of images, producing generally satisfactory results.

References

[1] Lucchese, L., Mitra, S.K.: Color image segmentation: A State-of-the-Art Survey. Proc. of the Indian National Science Academy (INSA-A) 67A(2), 207–221 (2001)

[2] Cheng, H.D., Jiang, X.H., Sun, Y., Wang, J.: Color image segmentation: advances and prospects. Pattern Recognition 34, 2259–2281 (2001)

[3] Freixenet, J., Muñoz, X., Raba, D., Martí, J., Cufí, X.: Yet Another Survey on Image Segmentation: Region and Boundary Information Integration. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2352, pp. 408–422. Springer, Heidelberg (2002)

[4] Sezgin, M., Sankur, B.: Survey over image thresholding techniques and quantitative performance evaluation. Electronic Imaging 13(1), 146–165 (2004)

[5] Wirjadi, O.: Image and video matting: a survey. Fraunhofer Institut für Techno und Wirtschaftsmathematik ITWM 2007, ISSN 1434-9973 Bericht 123 (2007)

[6] Zhang, H., Fritts, J.E., Goldman, S.A.: Image segmentation evaluation: A survey of unsupervised methods. In: CVIU, vol. 110, pp. 260–280 (2008)

[7] Shamir, A.: A survey on mesh segmentation techniques. Computer Graphics Forum 27(6), 1539–1556 (2008)

[8] Senthilkumaran, N., Rajesh, R.: Edge detection techniques for image segmentation: a survey of Soft Computing Approaches. Int. J. Recent Trends in Engineering 1(2), 250–254 (2009)

[9] Yang, Z., Chung, F.-L., Shitong, W.: Robust fuzzy clustering-based image segmentation. Applied Soft Computing 9, 80–84 (2009)

[10] Beucher, S., Lantuéjoul, C.: Use of watersheds in contour detection. In: Proc. Int. Workshop on Image Processing, Real-time Edge and Motion Detection/Estimation, Rennes, France, pp. 12–21 (1979)

[11] Vincent, L., Soille, P.: Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Trans. PAMI 13(6), 583–598 (1991)

[12] Roerdink, J.B.T.M., Meijster, A.: The watershed transform: definitions, algorithms and parallelization strategies. Fundamenta Informaticae 41, 187–228 (2001)

[13] Frucci, M.: Oversegmentation reduction by flooding regions and digging watershed lines. IJPRAI 20(1), 15–38 (2006)

[14] Frucci, M., Perner, P., Sanniti di Baja, G.: Case-based reasoning for image segmentation by watershed transformation. In: Case-based Reasoning on Images and Signals, vol. 73, pp. 319–352. Springer, Berlin (2007)

[15] Soille, P., Vogt, P.: Morphological segmentation of binary patterns. Pattern Recognition Letters 30(4), 456–459 (2009)

[16] Maulik, U.: Medical image segmentation using genetic algorithms. IEEE Trans. Information Technology in Biomedicine 13(2), 166–173 (2009)

[17] http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/

Supervised Learning Based Stereo Matching Using Neural Tree

Sanjeev Kumar[1], Asha Rani[2], Christian Micheloni[2], and Gian Luca Foresti[2]

[1] Department of Mathematics, IIT Roorkee, Roorkee 247667, India
[2] Department of Mathematics and Computer Science, University of Udine, Udine 33100, Italy

Abstract. In this paper, a supervised learning based approach is presented to classify tentative matches as inliers or outliers obtained from a pair of stereo images. A balanced neural tree (BNT) is adopted to perform the classification task. A set of tentative matches is obtained using speedup robust feature (SURF) matching and then feature vectors are extracted for all matches to classify them either as inliers or outliers. The BNT is trained using a set of tentative matches having ground-truth information, and then it is used for classifying other sets of tentative matches obtained from the different pairs of images. Several experiments have been performed to evaluate the performance of the proposed method.

Keywords: Neural Tree, Stereo Vision, Tentative Matches, Supervised Learning.

1 Introduction

One of the major challenges in wide baseline stereo images is the detection and removal of the outliers from a set of tentative matches due to its applicability in many vision applications such as construction of mosaic images, rectification of stereo images, change detection etc. Most of these applications require the estimation of a transformation matrix called homography. Such a homogarphy is estimated by minimizing a linear or nonlinear function for a given number of matching points. In this context, the nature of the homography estimation process is very sensitive with respect to false matches. Most of the image matching algorithms [1] contain three phases. In the first phase features are detected in both images in such a way that the detected features have similar appearance in different images (i.e., invariant [2]). Then, feature descriptors are computed as the signatures of the features. In the third phase the features of the first image are compared to the features of the second image. The comparison is performed using a suitable distance measure on the descriptors, and the tentative matches are ordered by similarity.

Scale invariant feature transform called as 'SIFT' [2] and speedup robust features (SURF) [3] based descriptors are very popular and effective approaches for obtaining the pairs of matching points between stereo images. However, when a

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 178–188, 2011.

wide baseline stereo system is used, the number of outliers becomes considerable. Therefore an extra phase is required to classify and remove such outliers from tentative matches. In classical approaches, these outliers are removed by fitting an affine or perspective transformation model to the top tentative matches. The model is then used to classify each tentative match as an inlier or outlier. All RANdom SAmple Consensus (RANSAC) based approaches [4,5] come under this category. Recently, an affine invariant model based approach has been proposed to classify tentative matches as inliers or outliers [6] with an advantage that it does not rely on the typical model generation as in RANSAC-based methods. These approaches are unsupervised in nature and not very effective when a large number of outliers are present in the respective set of tentative matches as in the case of wide baseline stereo images shown in Fig. 1. In such a scenario, the use of supervised learning based approaches is a more sensible choice in place of unsupervised techniques.

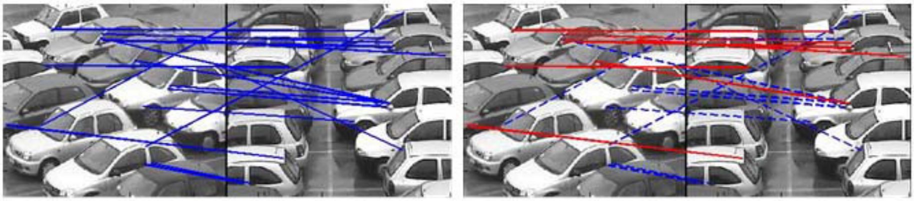


Fig. 1. Wide baseline stereo image matching. A set of tentative matches using SURF (left), Outlier detection using RANSAC (right), where blue dotted line represents detected outliers from the set of tentative matches.

In order to establish the point matching between a pair of stereo images, few supervised learning based methods have been proposed [7,8], where different machine learning tools like support vector machines (SVM) and artificial neural networks (ANN) have been used. These approaches are effective for performing the matching on the edges in stereo images, but at the same time have few drawbacks like the selection of an optimal architecture of the neural network for getting faster convergence with good accuracy. Moreover, these approaches are based on the matching of various edges between two images which makes them computationally expensive.

In this work, we are proposing a modified version of a neural tree [9] called balanced neural tree (BNT) for classifying inliers and outliers from a set of tentative matches (see Fig. 2). This BNT is composed by simple (single layer) perceptrons at its various nodes. Instead of local matching between images such as edge matching [8], we use a set of tentative matches obtained with SURF descriptors. For each tentative match in a set, a five dimensional feature vector (or pattern) is extracted and such a pattern is used as the input of the BNT. The BNT is trained in such a way that it achieves a balanced structure for better classification in less time. The proposed method has some advantages over existing

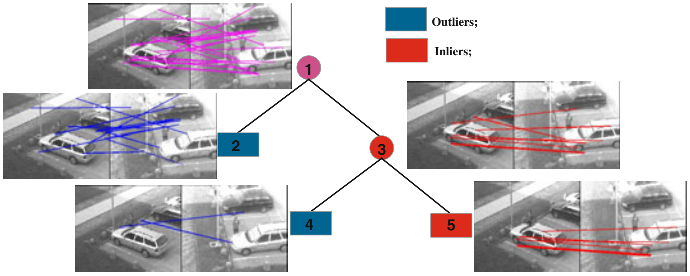


Fig. 2. An overview of the proposed framework: classification of tentative matches as inliers or outliers with neural tree's growth

unsupervised as well as supervised learning based approaches like its classification accuracy does not depend on the nature of data sets as in RANSAC and no need to decide an optimal architecture of network as in mutilayer perceptrons based approaches.

2 Features and Attributes Extraction

A pattern is generated corresponding to each tentative match based on the similarity between the descriptors and their respective local neighborhood in two images. Two different steps are required to obtain the necessary information for classifying a tentative match as inlier or outlier: (1) extraction of tentative matches from the pairs of stereo images, and (2) construction of feature vectors (or patterns) from these tentative matches.

The tentative matches are extracted using the SURF descriptor [3] due to its almost real time performance. The SURF detector is based on the Hessian matrix [10], but uses a very basic approximation like Laplacian-based detector. It relies on integral images to reduce the computation time and therefore called the 'Fast-Hessian' detector. The descriptor, on the other hand, describes a distribution of Haar-wavelet responses within the neighborhood of points of interest. Again, integral images are exploited for speed. Moreover, in SURF descriptors, only 64 dimensions are used for reducing the time of feature computation and matching, as well as increasing the robustness. The first step consists of fixing a reproducible orientation based on information from a circular region around the point of interest. Then, we construct a square region aligned to the selected orientation, and extract the SURF descriptor from it.

For each pair of descriptors belonging to a set of tentative matches, a total five different measures (the magnitude of the gradient vector, the direction of the gradient vector, the Laplacian, the variance and the mean) are computed. In this process, the gray levels of a central pixel (descriptor) and its eight immediate neighbors are used with the help of a 3×3 window. To find the gradient magnitude of the central pixel, we compare the gray level differences from the four pairs of opposite pixels in the 8-neighborhood, and the largest difference is taken as the gradient magnitude. The gradient direction of the central pixel is the direction out of the eight principal directions whose opposite pixels yield the largest gray level difference and also points in the direction which the pixel gray level is increasing. It is measured in degrees and a quantification followed by a normalization is performed as in [8]. The Laplacian is computed by using the corresponding Laplacian operator over the eight neighbors of the central pixel. The variance indicates the dispersion of the nine gray level values in the eight-neighborhood computed after a normalization process based on a linear regression.

Using the above described procedure, two different five dimensional vectors x_l and x_r are obtained corresponding to a given tentative match. Here, the components are the measures and the sub-indices l and r indicates that either vectors are corresponding to the left or right images, respectively. By taking the absolute difference of these two vectors $x = |x_l - x_r|$, we obtain the required feature vector or pattern x, whose components are corresponding to the differences for the module of the gradient vector, the direction of the gradient vector, the Laplacian, the variance and the mean, respectively.

3 Proposed Classifier: Balanced Neural Tree

Neural trees (NTs) have been developed to overcome the limitations of decision trees and neural networks based classification approaches. A neural tree grows during the learning phase and so it does not require any priori information about the network architecture (like number of hidden neurons or hidden layers). Moreover, it does not require any exhaustive search as used in training algorithms for decision trees. The first neural tree has been proposed in [11], whose internal nodes are represented by attribute tests and leaf nodes by perceptrons. Later, in [12], attribute tests at internal nodes are replaced by perceptrons. In some cases, a single layer perceptron can stuck in local extremals resulting in poor generalisation of the patterns which leads to a unbalanced or nonconvergent tree building process. In [9], a solution has been provided to this problem by introducing split nodes. Although split nodes assure the convergence of tree building process, but can generate a deep and unbalanced tree. Here, we propose a new tree architecture, called balanced neural tree (BNT), able to reduce the size of the tree with good classification accuracy. Two main novelties are proposed to achieve such a result: 1) balance the structure of the tree by substituting the current trained perceptron with a new perceptron if the current trained perceptron largely misclassifies the current training set into reduced number of classes, and 2) adopt a new criterion for the removal of tough training patterns that generate an over-fitting problem.

Learning algorithm of the proposed BNT is inherited from [9]. However, some novelties are introduced in order to overcome the above mentioned drawbacks. Let $S = \{(x_j, c_i) | j = 1, \ldots, n \wedge i \in [1, C]\}$ be the training set containing n number of k-dimensional x_j patterns belonging to a class c_i out of the possible C classes. The learning phase of the BNT is described in the Algorithm 1. Let S' be the local training set at the current node v, Q_v and Q_S be the queues holding the nodes and corresponding local training sets to be learned.

Algorithm 1. Learning phase of BNT classifier

```
Set $Q_v = \{v_0\}$ and $Q_S = \{S\}$
while ($Q_v$) do
    $v = Pop\ (Q_v)$ and $S' = Pop\ (Q_S)$
    Train Perceptron($v, S'$)
    ($Q_{\hat{v}}, Q_{\hat{S}}$)=Classify and remove tough patterns ($v, S'$)
    if $|Q_{\hat{S}}| = 1$ then
        Split node($v, S'$)
        Set $S_{\hat{v}} = \{v_L, v_R\}$ and $S_{\hat{S}} = \{S_L, S_R\}$
    end if
    if Perceptron unbalanced then
        Substitute perceptron
        ($Q_{\hat{v}}, Q_{\hat{S}}$)=Classify ($v, S'$)
    end if
    while $Q_{\hat{S}}$ do
        if $\tilde{S} = Pop\ (Q_{\hat{S}})$ is homogeneous then
            $\tilde{v} = Pop\ (Q_{\hat{v}})$ is set to leaf
        else
            $Push\ (Q_v, Pop\ (Q_{\hat{v}}))$ and $Push\ (Q_S, \tilde{S})$
        end if
    end while
end while
```

In the Algorithm 1, v_0 represents the root node, S is the training set at root node, $Q_{\hat{v}}$ and $Q_{\hat{S}}$ are the local queues to hold $\hat{v}$ and $\hat{S}$, respectively. The descriptions of the functions "Train perceptron", "Classify and remove tough patterns", "Split", and "Substitute Perceptron" mentioned in the above learning algorithm are explained below:

Train Perceptron(v, S')**.** It trains a single layer perceptron at node v on the training set S'. The perceptron learns until the error is not reducing any more for a given number of iterations. A trained perceptron generates $(o_1, ...o_c)$ activation values.

Classify and remove tough patterns(v, S')**.** It assigns the pattern to the class corresponding to the highest activation value and removes the patterns that are tough to classify. In other words, it divides S' in to $\hat{S} = \{S_1, ..., S_c\}$, $c \leq C$ subsets and generates next level of child nodes $\hat{v} = \{v_1, ...v_c\}$, $c \leq C$ corresponding to $\hat{S}$. Returns $(\hat{v}, \hat{S})$.

Split(v, S')**.** It divides the current training set into two parts S_L and S_R and generates corresponding child nodes v_L and v_R.

Substitute Perceptron. It substitutes an unbalanced perceptron with a new perceptron whose weights are initialized in such a way that the hyperplane generated passes through the barycentre of the training set.

The studied strategy bases its decision about which splitting criteria must be adopted on the basis of the classification errors. In particular, at each node the global classification error is computed as $E_t = 1 - (K_c/K_t)$, where K_c is the number of correctly classified patterns and K_t is the total number of patterns presented at the current node. In addition, to better localise the error among the classes, for each class a misclassification error is computed as $E_i = 1-(K_{c_i}/K_{t_i})$, where K_{c_i} is the number of correctly classified patterns of class i, and K_{t_i} is the total number of patterns classified as class i. During training, the criterion to judge whether a perceptron classification is acceptable is given as:

$$E_t > \frac{E_0}{m} \qquad \text{and} \qquad (E_{max} - E_{min}) > E_t \tag{1}$$

where m is the number of classes at the current node, $E_{max} = \max_i \{E_i\}$, $E_{min} = \min_i \{E_i\}$ and E_0 represents the initial error. If a perceptron is not able to separate the training set according to above mentioned criterion, then such a classification is not accepted and the corresponding perceptron is replaced with a specially designed perceptron that distribute the training set among all the classes equally. This kind of perceptron is generated by passing the splitting hyperplane through the barycenter of training set. The third kind of splitting is done when the perceptron fails to separate the patterns. For such a purpose a splitting rule [9] is considered to divide the data in two groups with almost equal cardinality. Concerning the second novelty of the proposed training algorithm, it has been noticed that even a well trained perceptron is not able to classify certain patterns. Such patterns are responsible for over-fitting in the training set. Such patterns are removed from the training set based on two facts, i.e., the probability of a pattern belonging to a class and the total classification error of the perceptron.

The classification probability is modelled by normalizing the activation values so that they form a distribution. The uncertainty between the two classes is given as $h_{ij} = |P(c_i|\mathbf{x}) - P(c_j|\mathbf{x})|$, where $P(c_i|\mathbf{x})$ and $P(c_j|\mathbf{x})$ are the probabilities of pattern $\mathbf{x}$, belonging to classes c_i and c_j respectively. Let $P(c_{max1}|\mathbf{x})$ and $P(c_{max2}|\mathbf{x})$ be the maximum and second maximum classification probabilities, thus representing the two most probable classes to which $\mathbf{x}$ should belong, then $h_{max} = |P(c_{max1}|\mathbf{x})$-$P(c_{max2}|\mathbf{x})|$ is defined to represent the uncertainty of the trained perceptron.

Concerning the reliability of the perceptron classification, studying the behaviors of the nodes with respect to their depth, suggested to define the reliability $R = 1/m^2$. If the current total error E_t is lower than the reliability factor R then the perceptron can be considered reliable.The criterion to decide whether a pattern has to be removed from the training set or not, is based on the following rule:

```
if $h_{max} < Th$ and $E_t < R$ and $\mathbf{x} \notin c_{max1}$ then
    Pattern $\mathbf{x}$ must be removed
else
    Pattern $\mathbf{x}$ is included in $TS_{max1}$
end if
```

Once the tree is build by learning the patterns, it can be used to classify the test patterns. The top to down traversal scheme [9] is used to classify the patterns. A pattern starts traversing the tree from the root node and traverse down until it reaches a leaf node. The next node to be considered on the path is decided by "winner-takes-all" rule. The class to be predicted is given by the leaf node reached by the pattern.

To perform the inlier and outlier classification on a set of tentative matches, BNT is trained on training data. In an online process, a set S having n pattern samples of tentative matches is classified into two classes inliers or outliers. The outputs of the system are two symbolic values $(+1, -1)$ each corresponding to one of the classes. During the classification process there are unambiguous and ambiguous matches, depending on whether a given left image segment corresponds to one and only one, or several right image segments, respectively. In any case, the decision about the correct match is made by choosing the pair with the higher activation value.

4 Experimental Results

Our aim is to check the performance of the proposed inliers-outliers classification method on different pairs of images. To do this, we have captured thirty pairs of stereo images of six different scenes. These image pairs are captured with a calibrated and wide baseline stereo cameras setup under varying illumination and baseline distance. Tentative matches are extracted from all image pairs using SURF descriptor followed by the earlier described process for obtaining respective feature vectors. All these feature vectors are put into six different sets (TM1-TM6) of tentative matches based on each scene. To check the performance of the proposed method, we have compared our classification results with different unsupervised (RANSAC [4] and affine invariant [6]) and supervised (multilayer perceptron [13] and AdaBoostM1 [14]) learning based methods. Here, employed multilayer perceptron has been composed by one hidden layer having four nodes and trained with a backpropagation learning algorithm.

Since we know the calibration parameters of the employed cameras, the ground-truth information have been extracted for the tentative matches (i.e., whether a match belong to inlier or outlier). The objective behind having the ground-truth is to know the percentage of correctly classified matches from tentative matches, i.e., evaluation of the accuracy of the proposed algorithm.

First, the set TM1 having total 457 tentative matches is used as a training set. In this set, classes (inlier or outlier) are assigned to each match and then this data is used for tree learning/building procedure. Once the tree is built, it is used to classify the tentative matches from other five sets (TM2-TM5) obtained with the images of different scene as well as illumination. This procedure has

Table 1. Classification accuracy using simple training strategy

Data Sets	Proposed BNT	Multilayer perceptron	AdaBoostM1 [14]	RANSAC
TM1	98.96	99.20	99.05	94.40
TM2	98.66	98.50	97.95	91.15
TM3	98.21	97.58	96.95	94.54
TM4	97.42	97.79	96.64	95.21
TM5	98.71	97.41	97.64	93.70
TM6	97.94	98.19	97.82	91.40

been repeated according to leave-one-out strategy, i.e. training on a data set and testing on the rest data sets. Finally, average of the classification accuracies of all data set obtained in five runs are taken. For a detailed quantitative analysis, these results showing the classification accuracy of the patterns belonging to these sets using above mentioned methods are given in Table 1 in terms of the percentage of correctly classified patterns.

From this table, it is clear that the proposed method perform better than the RANSAC for each data set (TM2-TM6). It is also worth to notice that it performed better than multilayer perceptron (MLP) in case of data sets TM2, TM3 and TM5, while for data sets TM1, TM4 and TM6, the performance of the MLP is better than our proposed method. However, the main advantage of the proposed method over MLP is that there is no need to decide the network architecture for getting optimal performance. It is also noticeable that the proposed method performed better than the AdaBoost classifier except the data set TM1, where the performance are quite close and comparable. For a visual representation of these results, Fig. 3 represents the two stereo-image pairs along with classification results obtained using the proposed and RANSAC algorithms.

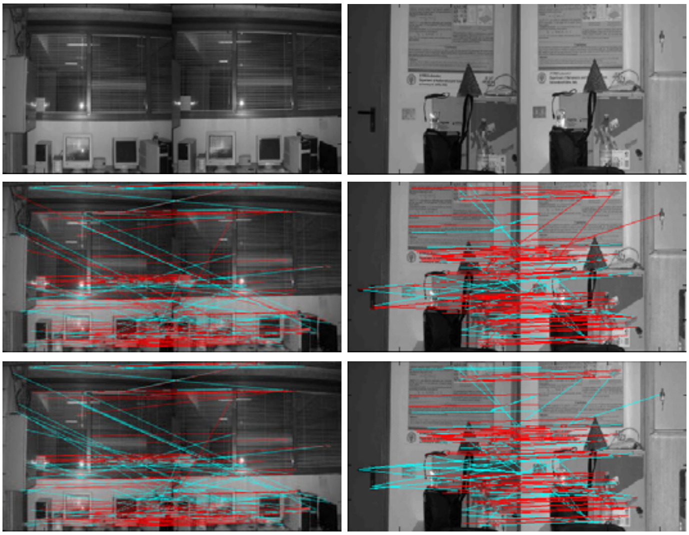


Fig. 3. Classification of tentative matches as inliers (red lines) and outliers (cyan lines). First row: two different stereo image pairs; second row: classification using RANSAC, and third row: classification using proposed BNT classifier.

Moreover, we have adopted an iterative process for the training of neural tree. First, the neural tree has been trained using data set TM1 and employed to obtain O(TM2) (where 'O' represents output) by classifying TM2. The classification results of TM2 has been combined with TM1, and again the neural tree has been trained using the combined (TM1+O(TM2)) training set. This

Table 2. Classification accuracy using iterative training strategy

Sets	Training Patterns	Testing Patterns	Proposed BNT	Multilayer perceptron
TM1	457	457	100.0	100.0
TM2	457	300	98.00	98.34
TM3	757	495	98.94	98.90
TM4	1252	520	99.60	99.65
TM5	1772	540	99.38	99.42
TM6	2312	614	99.38	99.50

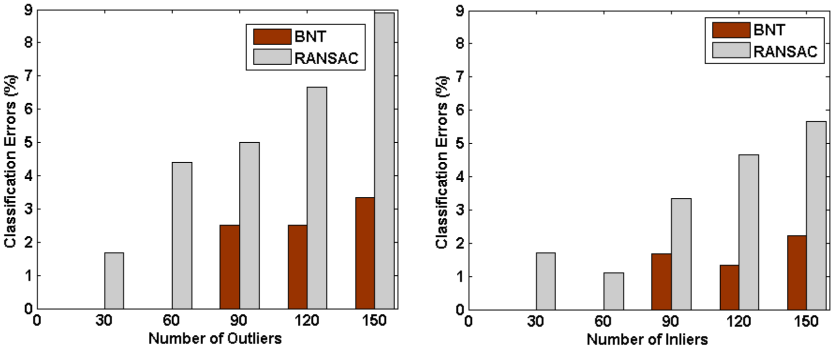


Fig. 4. Classification error in presence of different number of outliers/inliers in tentative matches

procedure has been repeated up to classification of TM6. The classification results obtained by adopting this iterative training strategy are given in Table 2. It is worth to notice that the result obtained with iterative training strategy is better than the earlier one in case of supervised learning algorithms. The reason behind this improvement is the better adoption of different factors like changes in the illumination and the scene for the training data of the neural tree. For observing the effect of the increasing numbers of outliers/inliers in the classification of tentative matches, we have conducted an experiment by considering different number of inliers or outliers in a set of tentative matches. In Fig. 4, a graph is shown between the classification error and increasing number of outliers, while keeping the fixed number of inliers. Initially, we considered 30 inliers (fixed) and 0 outliers in our set of tentative matches taken from data set TM6. This set has been classified by the BNT (trained on TM1) and RANSAC. Then, we have repeated this process five times by adding 30 outliers each time in the set of tentative matches and obtained the results. This experiments has been performed on five different samples and average of the classification errors are calculated. It has been noticed that the classification error increases very slowly in case of BNT, while is increases gradually in the case of RANSAC. A similar kind of results (see Fig. 4) by increasing number of inliers in the set of tentative matches while fixing the number of outliers as constant. Again, the variation in the error has been found very small in the case of BNT while it is more in the case of RANSAC.

It is also observed that most of the times only inliers are misclassified in the case of BNT classifier. It means that the inliers detected by the proposed method can be used for many sensitive operations in stereo vision like homography estimation, image rectification etc, as there is very minor chances for having an incorrect match. Here, it is also worth to mention that the proposed method gives a real time performance in terms of computational time.

5 Conclusions

We have proposed a new method to classify feature correspondences as inliers or outliers in stereo images. The proposed approach does not rely on the typical model generation and test approach used in RANSAC-based methods or in other unsupervised learning based classification approaches. A simple perceptron based neural tree has been employed to classify the tentative matches as inliers or outliers. The advantage of using neural tree over multilayer perceptron or other network architecture is that there is no need to decide an optimal network structure such as number of hidden layers or number of nodes in each hidden layer. It has been found that the proposed approach gives a very good result in different cases such as in different illuminations as well as scenes.

References

1. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. 27(10), 1615–1630 (2005)
2. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. Journal on Computer Vision 60, 91–110 (2004)
3. Baya, H., Essa, A., Tuytelaarsb, T., Gool, L.V.: Speed up robust features (surf). Computer Vision and Image Understanding 110(3), 346–359 (2008)
4. Torr, P., Zisserman, A.: Mlesac: a new robust estimator with application to estimating image geometry. Computer Vision and Image Understanding 78(1), 138–156 (2000)
5. Fischler, M.A., Bolles, R.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Comm. ACM 24(6), 381–395 (1981)
6. Fleck, D., Duric, Z.: Affine invariant-based classification of inliers and outliers for image matching. In: Kamel, M., Campilho, A. (eds.) ICIAR 2009. LNCS, vol. 5627, pp. 268–277. Springer, Heidelberg (2009)
7. Pajares, G., Cruz, J.: Local stereovision matching through the adaline neural network. Pattern Recognition Letters 22, 1457–1473 (2001)
8. Pajares, G., Cruz, J.: Stereovision matching through support vector machines. Pattern Recognition Letters 24, 2575–2583 (2003)
9. Foresti, G., Pieroni, G.: Exploiting neural trees in range image understanding. Pattern Recognit. Lett. 19 (9), 869–878 (1996)
10. Mikolajczyk, K., Schmid, C.: Indexing based on scale invariant interest points. In: Int. Conf. on Computer Vision, vol. 1, pp. 525–531 (2001)

11. Utgoff, P.E.: Perceptron tree: A case study in hybrid concept representation. Connection Science 1(4), 377–391 (1989)
12. Sankar, A., Mammone, R.: Neural Tree Networks. In: Neural Network: Theory and Application, pp. 281–302. Academic Press Professional, Inc., San Diego (1992)
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The weka data mining software: An update. SIGKDD Explorations 11(1) (2009)
14. Freund, Y., Schpire, R.: Experiments with a new boosting algorithm. In: Int. Conf. on Machine Learning, pp. 148–156. Morgan Kaufmann Pub. Inc., San Francisco (1996)

Pre-emptive Camera Activation for Video-Surveillance HCI

Niki Martinel, Christian Micheloni, and Claudio Piciarelli

Università degli Studi di Udine

Abstract. Video analytics has become a very important topic in computer vision. Many applications and different approaches have been proposed in different fields. This paper introduces a new information visualisation technique that aims to reduce the mental effort of security operators. A video analytics and a HCI module have been developed to reach the desired goal. Video analysis are exploited to compute possible trajectories used by the HCI module to pre-emptively activate cameras that will be probably interested by the motion of detected objects. The visualisation of most interesting views is provided to reduce the mental effort of human operators and allow them to follow the object of interest. Usability tests show the efficiency of the proposed solution.

1 Introduction

Video Surveillance Systems (VSS) are common in commercial, industrial, and now also in residential environment. The proliferation of VSSs and the increasing number of cameras per installation have brought new kind of problems. One of these is the usability of User Interfaces (UI).

Common video analytics tasks generally ask for significant mental efforts. Because of this, the limit of human attention is often overtaken and the real effectiveness, given the high number of video streams, tends to become a negative factor.

Today, this aspect is increasingly important because the primary goal of surveillance research is focused on the automatic understanding of the activities across wide areas [7]. When suspicious people have to be followed through multiple cameras users still experience great difficulties. For such a reason, VSSs must provide effective UIs such that the information represented by video streams and related data could be really understood by the end-users. User needs and application properties must be considered during the UIs development. Within last years many systems have been equipped with huge wall screens and/or some remote smaller displays. Anyhow, the required mental effort is not significantly decreased.

To follow objects between camera views users often employ a single monitor which dimensions are generally too small [3]. This kind of activity requires the user a huge mental effort and tends to become harder and harder if done for a long time. To sidestep such a problem, Iannizzotto et al. in [6] proposed a

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 189–198, 2011.

perceptual user interface that allows users interaction by means of gestures. In [5], Girgensohn et al. developed a Dynamic Object Tracking Systems that provides geographical cues [13] about the monitored environment. Morris and Trivedi in [9] and Bottoni et al. in [1] described similar solutions. According to these works new VSSs still have to strongly reduce the required mental effort.

The development of an effective and powerful information visualisation technique is the main goal of this work. The novel aspect of the paper is represented by the selection of proper video streams, their organisation and alternation. For such purposes, objects, i.e. camera views, are chosen and activated with a significantly difference: rather than displaying all available camera views, only most probable streams, i.e. those that will be interested by the object motion, are presented to the user. To determine most probable streams, the system must foresee the object trajectories and the cameras that best acquire such possible paths.

In [12], Qureshi and Terzopoulos describe how it is possible to activate different camera views in order to track a single object among different fields-of-view (FOV) that are geographically adjacent. Thus, merging gained knowledge about trajectory prediction and camera view selection and activation, the proposed work introduces a new way of showing visual information. The user interface developed is the result of an accurate process of camera view planning and selection. The main algorithm, according to space properties and a trajectory prediction tool, builds an activation plan to represent only those views that follow the predicted path of the object of interest.

Another novel aspect of this paper is related to the visualisation of geographical information about the monitored environment. Common desktop solutions, that made use of topographic maps, propose only a single topological representation of the environment. In addition, the proposed UI advantages from information visualisation studies conducted among mobile devices[2]. These exploit the introduction of the *detail plus overview* representation technique [14].

2 System Description

As shown in Fig.1 the architecture of the proposed VSS is organised in two main modules: a) a Video Analytics Module (VAM) and b) a Human-Computer Interface (HCI). The VAM processes the video streams generated by the cameras in order to analyse and identify events of interest [4] that should be provided to human operators together with useful information. For such a purpose a low level analysis module detects and recognises all the active objects in the environment. Then, the objects are tracked in order to provide temporal information about their activities. These are analysed by an event analysis module able to correlate the objects activities through time and space. Such an analysis is used by a trajectory estimator [11] that has the goal to foresee the trajectories of the objects of interest given their activities and past trajectories. Hence, this module path-plans the actions of interest such that the cameras can be opportunely tasked or redirected in order to improve the analysis capabilities. This is

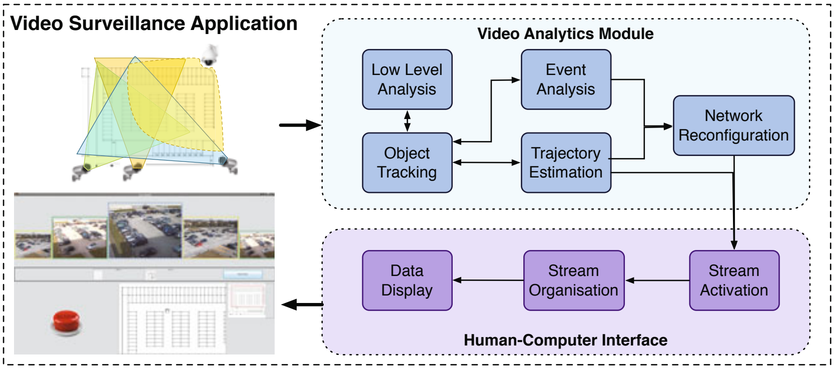


Fig. 1. Video Surveillance System

achieved by the network reconfiguration module [8]. The estimated trajectories and the new camera network configuration are given to the HCI module that has to optimise the cognitive capabilities of the operator. In this way, a first decision is taken about the most meaningful streams that have to be provided. This is done by considering the foreseen evolution of the environment. Once the most important streams have been decided a second module is in charge to determine how they must be organised on the user interface. Finally the streams are properly visualised on the UI together with useful information provided by the video analytics module.

3 Trajectory Analysis

The trajectory analysis module is fundamental to estimate the path that the object of interest most probably will take in the near feature. This module is based on Piciarelli and Foresti work [11]. Let $Tr = \{(x_1, y_1), \ldots, (x_n^t, y_n^t)\}$, where (x_j^t, y_j^t) is the position of the object expressed along the x^t and y^t axis of the map at time instant j, be the i-th trajectory detected by the trajectory extraction algorithm. The trajectories are grouped in clusters $C = \{(x_1^c, y_1^c, \sigma_1^2), \ldots, (x_m^c, y_m^c, \sigma_m^2)\}$ where σ_j^2 is the local variance of the cluster at time j that statistically represent the most probable trajectories inside the monitored environment. In order to associate the current trajectory T to an available cluster C the following distance measure is adopted

$$D(T, C) = \frac{1}{n} \sum_{i=1}^{n} min_j \left(\frac{dist((x_i^t, y_i^t), (x_j^c, y_j^c))}{\sigma_j} \right) \tag{1}$$

where *dist* is the Euclidean distance. Such a distance, representing the mean normalized distance of a trajectory point with the closest point of the cluster, is thresholded in order to associate it with a pre-computed cluster or to define a new one. Thus the matching process, as can be seen in Fig.2, allows to define a

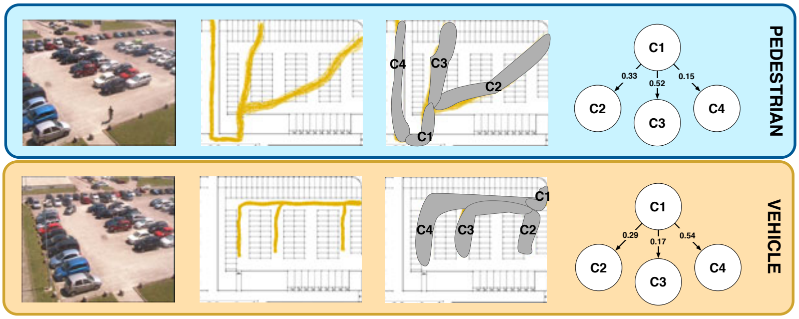


Fig. 2. Example of trajectory association and probability of future path

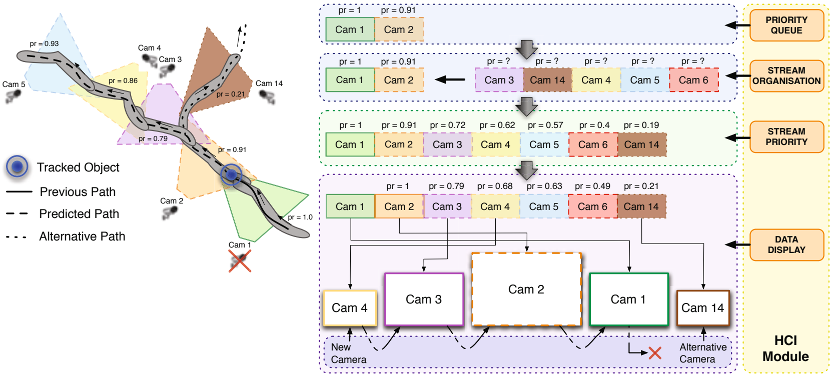


Fig. 3. HCI Module, example of stream organisation, activation and data display

probable path that the current object will follow. Matching the clusters positions with the cameras FOV allows to gain the probability that a sensor will acquire the object of interest in the near future. Thus, it is possible to have the activation probability of the cameras in order to correctly follow the object acquired by a camera.

4 HCI Module Definition

The HCI module is organised in three main components: a) the stream activation, b) the stream organisation and c) the data display.

4.1 Stream Activation

The stream activation makes active only those cameras that could be of interest for the operator. This HCI component analyses all available contents generated by the cameras and plans the hand-off between sensors [12] such that an object of interest could be continuously followed. The stream activation component provides an effective tool that is mainly based on the information provided by the trajectory estimation and network reconfiguration modules. In particular, the trajectory analysis provides the estimated path of the object. This is correlated with the fields-of-view computed by the network reconfiguration module. In this way, the stream activation foresees the cameras that will, most probably, be interested by the motion of the object of interest. Thus, such cameras are included in a priority queue that will allow to continuously maintain the selected object on the interface.

Camera views are inserted into the priority queue as follows. If Q is the priority queue and cam_i is the i-th camera such that $i \in \{1, ..., n\}$ where n represents the number of cameras, then cam_i is inserted in the priority queue if and only if

$$FOV_i \cap trajectory_j \neq \emptyset \tag{2}$$

where j is the j-th possible trajectory that the object could perform and FOV_i is the area on the ground covered by cam_i.

4.2 Stream Organisation

As show in Fig.3 this module has to re-weight all the streams that have been inserted in the priority queue. This is strictly necessary to allow a correct data interpretation. The goal is achieved by applying some organisation rules based on spatial relations between cameras views and object predicted trajectories. The algorithm is strongly dependent on the trajectories provided by the VAM module and by the stream activation component.

Streams that have been previously inserted into the priority queue are evaluated against the possible object trajectories by taking in account the geographical deployment of the sensors. Thus, according to the trajectory estimation component single view priority is calculated intersecting each trajectory cluster with each camera FOV. A priority value is then assigned to each camera.

The stream priority value is computed by traversing the predicted path tree (see Fig.2). Edges values $P(C_i, C_j)$, connecting the clusters C_i and C_j, represent the probability to move from cluster C_j to cluster C_i. Hence

$$P(C_i | C_j, C_{j-1} \ldots, C_k) = P(C_i, C_j) \prod_{l=j}^{k+1} P(C_l, C_{l-1}) \tag{3}$$

is the probability that the object will reach the cluster C_i through the path C_i, C_j, $C_{j-1}, \ldots, C_k$. The camera selected and covering the cluster C_i is assigned with a priority value equal to $P(C_i | C_j, C_{j-1} \ldots, C_k)$. The camera covering the current cluster is assigned with priority equal to 1. Once the priority values have been computed, the queue is sorted in order to have higher priority cameras on top.

4.3 Data Display

The described UI introduces a new way of displaying multiple video streams in a single device by using a strictly user-centred development process.

The introduced data display module has two main novelties that have been developed to improve the operator capabilities: a) the video streams area and b) the map area. The main UI component is represented by the video area that is based on the two previously described components. Such a UI element is able to display only those cameras that better support the operator activities necessary to monitor the selected object. In addition, to provide an effective user interface the map of the area is also displayed. On such a map, sensors and objects positions are displayed.

The data display module is object centred since the visualisation is independent of the number of cameras in the network. It depends only on the objects of interest tracked by the system. In this way, when tracking an object, the number of cameras interested by the task is fixed. If more objects of interest are present into the monitored area the operator is able to follow one of these just switching, through a tabbed pan, the active visualisation. The number of objects of interest that are associated to a single operator is limited.

Video Streams Area. The video streams area represents the most important UI component of this work. The visualisation techniques adopted have been tailored according to the results obtained from different evaluations conducted among a set of preselected users. As previously described video streams are arranged inside the priority queue following some basic rules linked to the spatial properties between cameras and the predicted trajectory of the tracked object (see section 4.2).

Following the priority queue, the most important view, i.e. the stream of the sensor whose FOV best shows the object (and has the highest priority), is always displayed at the centre of the streams area. The streams that came after the most important camera will be displayed at the side of the main camera view according to the movement of the object. The previous camera with the highest priority will be shown at the other side. So, thanks to this the operator can clearly see which is the previous and the next camera view that will be used to follow the object itself. Finally the camera view that has been assigned the highest priority and is not related to the main predicted path, i.e. the most probable alternative path, will be displayed at the extreme side of the main camera view such that if the tracked object does not follow the main predicted path the operator can still see it.

Assuming that the object is moving from right to left, the next view will be presented to the operator at the left of the current main view (see Fig.3). In this way, as the main view changes, the next predicted stream will be moved to the centre of the visualisation area. Adopting this technique the object moving out of the main camera view will appear in the next stream represented in the direction of the object itself.

The video stream area introduces another visualisation cue that aims to better support operator tasks. Such a cue is described by the size of a single video stream representation that has the goal to explain the relative importance of each camera view. Adopting the described view selection technique, together with the view size cue, elements importance could decay exponentially over time so that video streams stay in view for some time after they become less important. In some cases, as the object move along its path, it could be possible that the camera selection and representation change too quickly inside the UI. This way, the user could get confused. So, to solve this possible problem the user interface introduces content animation. As long as the object follows the predicted path, camera views gradually slide in the opposite direction respect to the movement vector of the object. Old selected views are scaled down and animated out of the UI.

Other important aspects that have been introduced in the UI are represented by colours and by depiction techniques adopted to differentiate camera views. The UI introduces a colour-coded and a drawing style technique to discern camera views and to relate them to the represented objects in the map area. These UI features have been adopted to best fit all user needs and to reduce the operators mental effort. The main goal has been achieved using a color-coded technique (which complies with colour-blind people) and introducing a different representative style for the main active sensor. Hence, all users could clearly distinguish between environment sensors and in particular can immediately recognise main active camera.

Although providing only stream information could fit some operator requirements, in some cases it might be useful to have geographical information about the area and spatial cues about the identified objects that are followed by ambient sensors. Using a topological representation that cooperates with the previously described techniques is probably the best way that can lead to better usability results.

Map Area. In many VSSs personnel is often required to monitor many video streams and the introduction of a map component could help them and improve their ability to follow objects behaviour between different camera views. Usually, the map, that is integrated in this kind of user interfaces, displays the location of each camera, its FOV and the objects being tracked. Moreover the map can sometimes be panned and zoomed for retrieving more accurate details.

In this work three main novel aspect have been introduced on map visualisation. The first is represented by the colour-coded technique previously analysed (see section 4.3). The second is given by the use, for each camera, of a colour-coded shaded area representing the sensor FOV. Finally, taking care of all commonly adopted techniques, the work proposed through this paper also introduces a novel interaction paradigm that is usually employed in mobile devices applications: the *detail plus overview* technique. This powerful information visualisation technique significantly improve user ability to find and search objects inside the monitored environment. Adding a small representation of the topological area map allow operators to interact with this user interface element and easily navigate the environment maintaining a fixed view about the whole area. Furthermore, if the user has zoomed the map view to retrieve more accurate details about the tracked object path, and this is now outside the current view of the map, the user can continue to follow the object through the smaller representation. It could also pan the magnified view to update the focus on the new target position.

5 Experimental Results

As user-centred design principles pointed out, it's not possible to evaluate a project just executing some test on a single prototype. Thus, the work described has been conceived and developed using an iterative process that has the goal to produce the best interaction between the operator and the machine. This process has led to the development of four different prototypes that have been evaluated using empirical and non-empirical methods.

First of all some basic information about classes of users, context of use and application scenarios have been identified. This initial step drives the process to the second evaluation stage that is composed by the retrieving process. This has to identify the using cases and the user basic knowledge necessary to execute these evaluations. Then, evaluations tests have been executed and results analysed to identify and fix every single interaction problem.

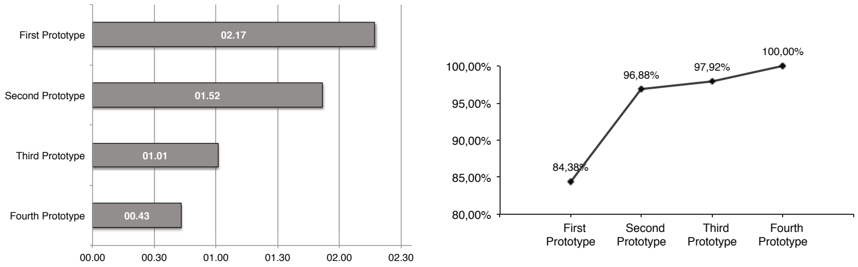
Empirical evaluations have been conducted among a set of about forty pre-identified end-users. They have been asked to execute six basic activities like viewing a past event, follow the tracked object and so on. Test sessions have then been executed in a controlled environment where the researcher maintains a detached behaviour and intervenes only when the end-user isn't able to reach the requested result or has some question about the UI elements behaviour.

To get some quantitative evaluation of the designed and developed UI some indexes have been defined. The success rate index (SR) aims to show how much efficient is the user interface. It is given by:

$$SR = \frac{n_s}{n_t} \tag{4}$$

where n_s is the number of correct-end test and n_t is the total amount of conducted evaluations. The information that arise from the SR index can then be used to gain an overview about the efficiency and clearness of the UI.

To get a more clear and accurate point of view, on each UI, single results obtained from the evaluation process have been extracted. Such information has been taken in account to precisely found what are the problems related to single human-machine task.



(a) (b)

Fig. 4. Results obtained from experimental results. (a) Average Execution Time Index (b) Success Rate Index.

Similarly to the SR index the average execution time (AET) has been calculated to gain some information about the user interface efficiency. The AET is computed as

$$AET = \frac{\sum_{i=0}^{n_t} T_i}{n_t} \tag{5}$$

where T_i represents the i-th execution time needed to complete a single proposed task. This index has been used to represent how much time a single user needs to reach the given goal (user failure has been taken into account as well). In particular, data obtained from this index is as important as the data collected from the SR index because timing (especially in computer vision tasks) is much significant. So, analysing data it was possible to identify which were the HCI tasks that require a specific amount of time to be completed. During the design process, if a given task required too much time the UI elements involved in that process had to be reviewed.

Non-empirical evaluations have been executed with the direct support of four HCI experts that try to complete the predefined tasks and thus identify which problem an end-user could find. In particular during this kind of evaluation have been used two commonly adopted techniques: a) the heuristic evaluation [10] and b) the cognitive walkthrough [15]. In both cases results obtained from evaluations have been crossed with empirical evaluation data to get the best results as possible.

As shown in Fig.4 results obtained demonstrate the effectiveness of a user centred design research process that directly involve end-users in the evaluation task. The AET index that has been evaluated against each prototype shows a constant reduction of the time necessary to complete a requested task. It shows very interesting values and demonstrates the efficiency of the fourth prototype that average require less than a minute to complete a single task. The SR index describes the relevance of the adopted approach too and shows once again the clearness of the last evaluated prototype (i.e. 100% of successful tests).

6 Conclusions

In this paper a novel information visualisation approach for VSSs has been described. A VAM has been developed to identify and predict the path of object of interest. Tracking data is used to evaluate object trajectories and determine camera configurations. A HCI module has been used to select, organise and show best streams to keep the object inside the UI. Obtained results have shown that the adopted information visualisation technique is very efficient and leads to a mental effort reduction for end-users.

References

1. Bottoni, P., De Marsico, M., Levialdi, S., Ottieri, G., Pierro, M., Quaresima, D.: A dynamic environment for video surveillance. In: Gross, T., Gulliksen, J., Kotzé, P., Oestreicher, L., Palanque, P., Prates, R.O., Winckler, M. (eds.) INTERACT 2009, Part 2. LNCS, vol. 5727, pp. 892–895. Springer, Heidelberg (2009)

2. Chittaro, L.: Visualizing information on mobile devices. IEEE Computer 39(3), 34–39 (2006)
3. Colineau, N., Phalip, J., Lampert, A.: The delivery of multimedia presentations in a graphical user interface environment. In: 11th International Conference on Intelligent User Interface, Sydney, Australia, pp. 279–281 (February 2006)
4. Foresti, G., Micheloni, C., Piciarelli, C.: Detecting moving people in video streams. Pattern Recognition Letters 26, 2232–2243 (2005)
5. Girgensohn, A., Kimber, D., Vaughan, J., Yang, T., Shipman, F., Turner, T., Rieffel, E., Wilcox, L., Chen, F., Dunnigan, T.: Dots: Support for effective video surveillance. In: Proceedings of the 15th International Conference on Multimedia, Augsburg, Germany, pp. 423–432 (September 2007)
6. Iannizzotto, G., Costanzo, C., Rosa, F.L., Lanzafame, P.: A multimodal perceptual user interface for video-surveillance environments. In: 7th International Conference on Multimodal Interfaces, pp. 45–52 (October 2005)
7. Lee, L., Romando, R., Stein, G.: Monitoring activities from multiple video streams: Establishing a common coordinate frame. IEEE Transactions on Pattern Analysis And Machine Intelligence 22(8), 758–767 (2000)
8. Micheloni, C., Rinner, B., Foresti, G.L.: Video analysis in pan-tilt-zoom camera networks. IEEE Signal Processing Magazine 27(5), 78–90 (2010)
9. Morris, B.T., Trivedi, M.M.: Contextual activity visualization from long-term video observations. IEEE Intelligent Systems 25(3), 50–62 (2010)
10. Nielsen, J.: Usability inspection methods. In: Conference Companion on Human Factors in Computing Systems, CHI 1994, pp. 413–414. ACM, New York (1994)
11. Piciarelli, C., Foresti, G.: Online trajectory clustering for anomalous event detection. Pattern Recognition Letters 27, 1835–1842 (2006)
12. Qureshi, F.Z., Terzopoulos, D.: Planning ahead for ptz camera assignment and handoff. In: Third ACM/IEEE International Conference on Distributed Smart Cameras 2009, Como, Italy, pp. 1–8 (August-September 2009)
13. Rieffel, E.G., Girgensohn, A., Kimber, D., Chen, T., Liu, Q.: Geometric tools for multicamera surveillance systems. In: First ACM/IEEE International Conference on Distributed Smart Cameras, pp. 132–139. ACM/IEEE (2007)
14. Spence, R.: Information Visualization. Addison Wesley, Harlow (2000)
15. Wharton, C., Rieman, J., Lewis, C., Polson, P.: Usability Inspection Methods, 1st edn., ch. 5, pp. 105–140. Wiley, Chichester (1994)

Space-Time Zernike Moments and Pyramid Kernel Descriptors for Action Classification

Luca Costantini[2], Lorenzo Seidenari[1], Giuseppe Serra[1], Licia Capodiferro[2], and Alberto Del Bimbo[1]

[1] Media Integration and Communication Center, University of Florence, Italy
{delbimbo,seidenari,serra}@dsi.unifi.it
[2] Fondazione Ugo Bordoni, Roma, Italy
{lcostantini,lcapodiferro}@fub.it

Abstract. Action recognition in videos is a relevant and challenging task of automatic semantic video analysis. Most successful approaches exploit local space-time descriptors. These descriptors are usually carefully engineered in order to obtain feature invariance to photometric and geometric variations. The main drawback of space-time descriptors is high dimensionality and efficiency. In this paper we propose a novel descriptor based on 3D Zernike moments computed for space-time patches. Moments are by construction not redundant and therefore optimal for compactness. Given the hierarchical structure of our descriptor we propose a novel similarity procedure that exploits this structure comparing features as pyramids. The approach is tested on a public dataset and compared with state-of-the art descriptors.

Keywords: video annotation, action classification, Zernike moments.

1 Introduction and Related Works

Human behavior recognition is a challenging computer vision task that have recently attracted wide research effort; this is mainly due to the need of automatic semantic analysis of video data in several application fields such as intelligent video-surveillance systems and digital libraries. In video surveillance it is often the case that human operators are simply not able to attentively observe a large amount of screens in parallel; moreover in forensics, retrieval of video footage containing well defined human actions is invaluable.

Several techniques have been developed in the recent years mainly based on the use of local descriptions of the imagery. Following the success of SIFT [1] in object and scene recognition and classification [2], several space-time extensions of the local patch descriptors have been proposed. Similarly to local image features [3,4] space-time features are localized through a detection step and then computed on the extracted patches; videos are represented as a collection of descriptors. Space-time descriptors represent the appearance and the motion of a local region and are engineered in order to retain invariance to geometric and

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 199–208, 2011.

photometric transformations. Laptev *et al.* [5] defined a descriptor as a concatenation of histograms of oriented 2D gradients and histograms of optical flow. In order to reduce the computation burden an extension of SURF have been presented in [6]. Scovanner *et al.* [7] extended the SIFT to three-dimensional gradients normalizing 3D orientations bins by the respective solid angle in order to cope with the issue of the uneven quantization of solid angles in a sphere. To solve this issue Kläser *et al.* [8] proposed to exploit 3D pixel gradients developing a technique based on Platonic solids. Finally Ballan *et al.* [9] developed an efficient descriptor decorrelating the spatial and temporal components and creating separated histograms of 3D gradient orientations. However, all of these descriptors are extremely high-dimensional and often retain redundant information.

In the same time, researchers have exploited moments and invariant moments in pattern recognition [10]. Moments are scalar quantities used to characterize a function and to capture its significant features and they have been widely used for hundreds of years in statistics for description of the shape of a probability density function. Moments and in particular Zernike moments are a common choice in shape representation [11]. Zernike moments have been also proposed in action recognition as holistic features in [12] to describe the human silhouettes.

Despite the fact that feature matching is an important step in the recognition process few works have analysed it. Lowe [1] showed that in order to retrieve meaningful patches it is necessary to look at the distances of the second nearest neighbour. More recently Bo *et al.* [13] provided a kernel view of the matching procedure between patches. Their work formulates the problem of similarity measurement between image patches as a definition of kernels between patches. Since these kernels are valid Mercer kernels it is straightforward to combine or plug them into kernelized algorithms.

In this paper we propose a new method for classification of human actions based on an extension of the Zernike moments to the spatio-temporal domain. Furthermore, we propose a kernel suitable for matching descriptors that can be hierarchically decomposed in order to obtain a multiple resolution representation. This kernel is inspired by multi-resolution matching of sets of features [14,15], but instead of matching sets of features we match single space-time patches at multiple resolutions. To the best of our knowledge 3D Zernike moments have never been used as local space-time features and the pyramid matching scheme has never been used to define kernels between single features but only to match sets of features. Experimental results on KTH dataset shows that our system presents a low computational time maintaining comparable performance with respect to the state-of-the-art. The rest of the paper is organized as follows. The generalization of the Zernike moments to the three dimensions is presented in the next section. The Pyramid Kernel Descriptors are introduced in Sect. 3. The techniques for action representation and classification are presented in Sect. 4. Experimental results on the standard KTH dataset are discussed in Sect. 5. Finally, conclusions are drawn in Sect. 6

2 Space-Time Zernike Moments

We first describe the formulation of the Zernike moments in two dimensions, and then introduce the generalization to the space-temporal domain. Let $\mathbf{x} = [x_1, x_2]$ be the Cartesian coordinates in the real plane $\mathbb{R}^2$. Zernike polynomials are a set of orthogonal functions within the unit disk composed by a radial profile R_{nm} and a harmonic angular profile $H_m(\vartheta)$ defined as follows

$$V_{nm}(\rho, \vartheta) = R_{nm}(\rho) \cdot H_m(\vartheta) \tag{1}$$

where $\rho = \sqrt{x_1^2 + x_2^2}$, $\vartheta = \tan^{-1}\left(\frac{x_2}{x_1}\right)$, $H_m(\vartheta) = e^{im\vartheta}$ and

$$R_{nm}(\rho) = \begin{cases} \sum\limits_{s=0}^{(n-|m|)/2} \frac{(-1)^s (n-s)! \rho^{n-2s}}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!} & \text{for } n-|m| \text{ even} \\ 0 & \text{for } n-|m| \text{ odd} \end{cases}. \tag{2}$$

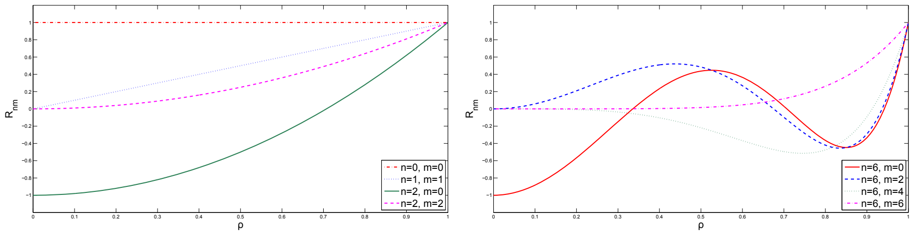
The index n is named “order” and is a non-negative integer, and m is called “repetition” and it is an integer such that $n - |m|$ is even and non-negative. In Fig. 1 some examples of the radial profile R_{nm} are shown. Both the Zernike polynomials and the radial profile $R_{nm}(\rho)$ satisfy the orthogonal condition

$$\int_0^{2\pi}\int_0^1 V^*_{nm}(\rho, \vartheta)\, V_{n'm'}(\rho, \vartheta)\, \rho d\rho d\vartheta = \frac{\pi}{n+1}\delta_{nn'}\delta_{mm'} \tag{3}$$

and

$$\int_0^1 R_{nm}(\rho)\, R_{n'm'}(\rho)\, \rho d\rho = \frac{1}{2(n+1)}\delta_{nn'}\delta_{mm'} \tag{4}$$

where δ indicates the Kronecker delta. Zernike polynomials are widely used to compute the Zernike moments [16,17].



(a)

(b)

Fig. 1. a) Radial profile up to the 2^{nd} order; b) Radial profile for the 6^{nd} order

Let $f(\mathbf{x})$ be any continuous function, the Zernike moments are

$$A_{nm}(\mathbf{x_0}) = \frac{n+1}{\pi} \int\int_{\|\mathbf{x}-\mathbf{x_0}\| \leq 1} f(\mathbf{x})\, V^*_{nm}(\mathbf{x} - \mathbf{x_0})\, dx_1 dx_2 \tag{5}$$

where $\mathbf{x_0}$ denotes the point where the unit disk is centered. In this work we are interested in the computation of the Zernike moments for functions as $f : \mathbb{R}^3 \mapsto \mathbb{R}$ where the third dimension is the time. To get the 3D Zernike polynomials [18,19], the harmonic angular profile is substituted by the spherical harmonic functions

$$Y_m^l(\vartheta, \varphi) = N_m^l P_m^l(\cos\vartheta)\, e^{il\varphi} \tag{6}$$

where P_m^l denotes the Legendre function and N_m^l is a normalization factor

$$N_m^l = \sqrt{\frac{2m+1}{4\pi} \frac{(m-l)!}{(m+l)!}}. \tag{7}$$

The spherical harmonic functions up to the 3^{rd} order are shown in Fig. 2. In this case, given an order n, we use only the values of $m \geq 0$, and the index l is an integer such as $-m \leq l \leq m$. Then, the 3D Zernike polynomials are defined in spherical coordinates as follows

$$V_{nm}^l(\rho, \vartheta, \varphi) = R_{nm}(\rho) \cdot Y_m^l(\vartheta, \varphi) \tag{8}$$

and they satisfy the orthogonal condition within the unit sphere

$$\int_0^1 \int_0^\pi \int_0^{2\pi} \left[V_{nm}^l(\rho, \vartheta, \varphi)\right]^* V_{n'm'}^{l'}(\rho, \vartheta, \varphi) \sin(\vartheta)\, d\vartheta d\varphi d\rho = \delta_{nn'} \delta_{mm'} \delta^{ll'}. \tag{9}$$

Let $\boldsymbol{\xi} = [\mathbf{x}, t]$ be the generic point in the real plane $\mathbb{R}^2$ at the time t, the 3D Zernike moments are

$$A_{nm}^l(\boldsymbol{\xi_0}) = \frac{3}{4\pi} \int\limits_{\|\boldsymbol{\xi} - \boldsymbol{\xi_0} \leq 1\|} f(\boldsymbol{\xi}) \left[V_{nm}^l\left(\frac{\boldsymbol{\xi} - \boldsymbol{\xi_0}}{\boldsymbol{\sigma}}\right)\right]^* d\boldsymbol{\xi} \tag{10}$$

where $\boldsymbol{\xi_0}$ is the point where the unit sphere is centered, and $\boldsymbol{\sigma}$ tunes the size in pixel of the unit sphere for each coordinate. This $\boldsymbol{\sigma}$ is necessary because the patches, that we need to describe by using the 3D Zernike moments, can have different sizes in space and time. We use these space-time Zernike moments as descriptors for the local patches. The orthogonal condition (see Eq. 9) ensures that there is no redundant information in the descriptor allowing to have a compact representation of the local feature. Fig. 3 shows that we can obtain a rough but representative reconstruction of space-time cuboid from the 3D Zernike moments. In particular, we exploit the phase of these complex moments since from preliminary experiments proved to be more effective.

3 Pyramid Kernel Descriptors

We introduce a descriptor matching kernel inspired by multi-resolution matching of sets of features[15,14]; Grauman and Darrel [15] proposed the Pyramid

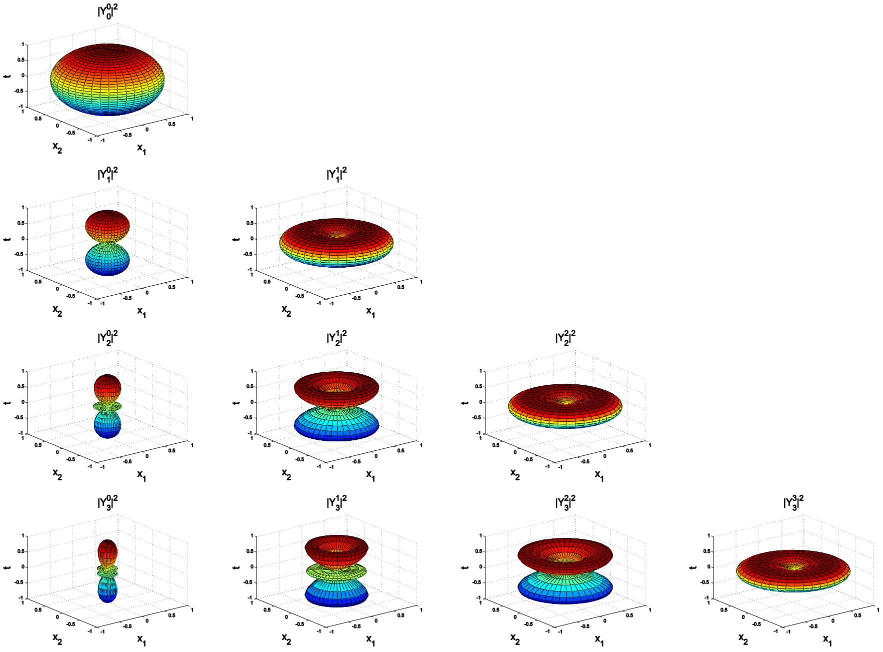


Fig. 2. Spherical harmonic functions up to the 3^{rd} order

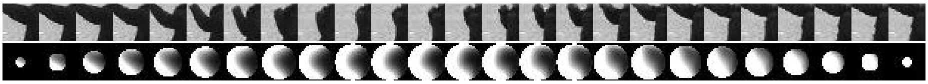


Fig. 3. Frames of a cuboid (top). Reconstructed cuboid from complex 3D Zernike moments up to the 6^{th} order (bottom).

Matching kernel to find an approximate correspondence between two sets of features points. Informally, their method takes a weighted sum of the number of matches that occur at each level of resolution, which are defined by placing a sequence of increasingly coarser grids over the features space. At any resolution, two feature points match if they fall into the same cell of the grid; number of matches computed at finer resolution are weighted more than those at coarser resolution. Later, Lazebnik *et al.* [14] introduced the Spatial Pyramid Matching kernel that work by partitioning the image into increasingly fine sub-regions and computing histograms of local features found inside each sub-regions.

Differently from these approaches our idea is to adapt the pyramid scheme for computing the similarity between two descriptor points. This allows to compute the similarity between two descriptors at multiple resolutions, exploiting a more distinctive representation when available and discarding it when at higher resolutions becomes noisy. We call our proposed approach "Pyramid Kernel Descriptors" because feature points are matched considering the descriptors as a multi-resolution set.

We consider a set of space-time interest points $X = \{\xi_1, \dots \xi_s\}$ and their descriptors $D = \{d_1, \dots, d_s\}$, where each descriptor can be organized in p sets $\{s^1, \dots, s^p\}$ hierarchically ordered. The pyramid kernel between d_i and d_j is defined as a weighted sum of the similarities of sets found at each level of the pyramid:

$$K(d_i, d_j) = \sum_{k=0}^{p} w_k k_c(s_i^k, s_j^k) \tag{11}$$

where w_k is the weight and $k_c(s_i^k, s_j^k)$ is a kernel to compute similarity between s_i^k and s_j^k. The similarity found at each level in the pyramid is weighted according to the description resolution: similarities made at a finer resolution, where features are most distinct, are weighted more than those found at a coarser level. Thus, if the p sets are arranged in ascending order the weight at level k can be defined as $w_k = 2^{k-p}$. If k_c is a valid kernel, our proposed kernel is a valid Mercer kernel for the closure property of kernels since it is a weighted sum of valid kernels. As described in sect. 2, our description based on space-time Zernike moments have a pyramid structure defined by the orders. In fact, lower order moments describe low frequencies of each cuboid while higher order moments encode higher frequencies. We define s^k as the concatenation of the phases of the complex Zernike moments for the first k orders: $s^k = \big(\arg(A_{00}^0), \dots, \arg(A_{km}^l)\big)$, where m and l are set according to Sect. 2. We use a normalized scalar product: $k_c(s_i^k, s_j^k) = \frac{s_i^k \cdot s_j^k}{\|s_i^k\| \|s_j^k\|}$, as a kernel between s_i^k and s_j^k, which is a valid Mercer kernel. Note that we normalize the scalar product computed at each level in order to have comparable values in the final sum.

For example, if we use a two level pyramid kernel descriptor then $s_0 = \big(\arg(A_{00}^0)\big)$, $s_1 = \big(\arg(A_{00}^0), \arg(A_{11}^{-1}), \arg(A_{11}^0), \arg(A_{11}^1)\big)$ and the corresponding weights are $w_0 = 1$ and $w_1 = \frac{1}{2}$. The final kernel between two space-time Zernike descriptors d_i, d_j computed up to the n^{th} order is:

$$K(d_i, d_j) = \sum_{k=0}^{n} 2^{k-n} \frac{s_i^k \cdot s_j^k}{\|s_i^k\| \|s_j^k\|}. \tag{12}$$

4 Action Classification

We represent an action as a bag of space-time interest points detected by an adaptation of the detector proposed by Dollár *et al.* [20]. This detector applies two separate linear filters to spatial and temporal dimensions, respectively. The response function has the form:

$$R = \Big(I * g_\sigma * h_{ev}\Big)^2 + \Big(I * g_\sigma * h_{od}\Big)^2 \tag{13}$$

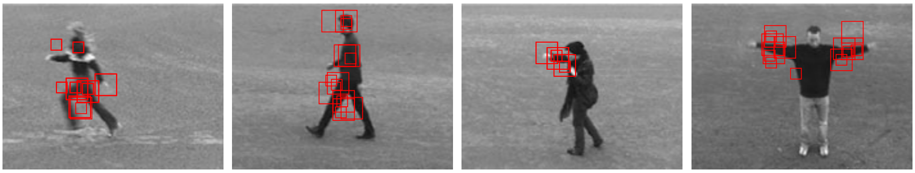


Fig. 4. Examples of space-time interest points extracted at multiple scales for different actions. Clips are taken from the KTH dataset: running, walking, boxing and hand-waving.

where $I(x, y, t)$ is a sequence of images over time, $g_\sigma(x, y)$ is the spatial Gaussian filter with kernel σ, h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied along the time dimension. They are defined as $h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$ and $h_{od}(t; \tau, \omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$, where $\omega = 4/\tau$. The interest points are detected at locations where the response is locally maximum typically corresponding to the temporal intensity changes. In order to cope with spatial and temporal scale variations we extract features at multiple scales. Usually these locations correspond to human body limbs involved in the execution of an action as can be seen in Fig. 4.

Each point is described using Space-time Zernike moments and then a nearest-neighbor classifier based on the concept of instance-to-class similarity [21] is used for action categorization. We choose not to employ descriptor codebooks (as in bag-of-words approaches) in order to better evaluate the effectiveness of our descriptor alone.

The instance-to-class nearest-neighbor classifier estimates the class posterior probability given the query video clip with a non-parametric density estimation based on local Parzen windows centered on descriptors belonging to the class. In [21] authors have shown that formulations based on more than one nearest neighbor per query descriptor do not significantly outperforms the simpler 1-NN formulation. Given this evidence, the implementation of this simple but effective classifier boils down to obtaining the most similar descriptor from the database for each feature extracted in a query clip (generally based on Euclidean distance between descriptors) and accumulating a vote for the class to which the database descriptor belongs to. The class with more votes is associated to the query clip. Instead of using Euclidean distance, we use our pyramid kernel descriptors (Sect. 3) to select the most similar descriptors which have, for each feature, the maximum kernel values.

5 Experimental Results

We tested our approach on the KTH action dataset containing six actions (walking, running, jogging, hand-clapping, hand-waving, boxing) performed several times by 25 actors under four different scenarios of illumination, appearance and scale change. The dataset contains 2391 video sequences with resolution of 160×120 pixel.

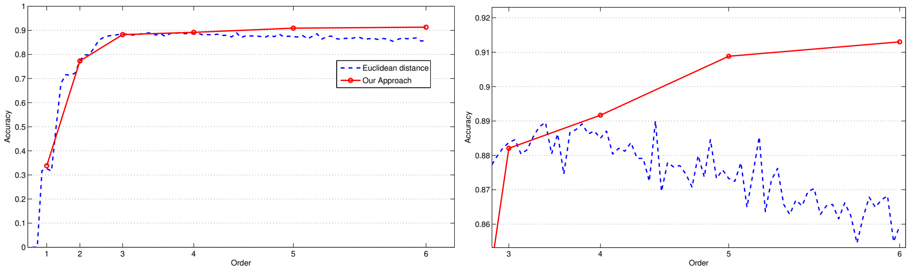
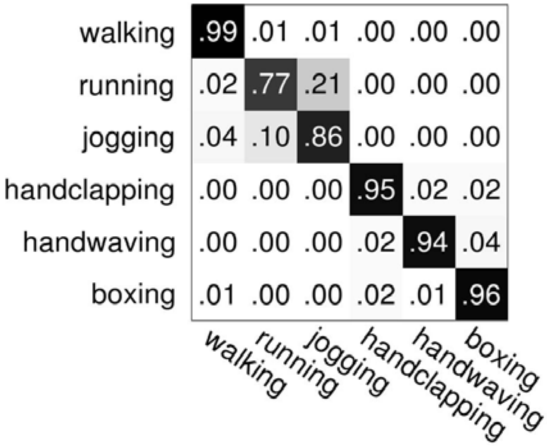


Fig. 5. Comparison of the two similarity techniques; right) detail showing the effect of pyramid matching descriptors on high order moments



	walking	running	jogging	handclapping	handwaving	boxing
walking	.99	.01	.01	.00	.00	.00
running	.02	.77	.21	.00	.00	.00
jogging	.04	.10	.86	.00	.00	.00
handclapping	.00	.00	.00	.95	.02	.02
handwaving	.00	.00	.00	.02	.94	.04
boxing	.01	.00	.00	.02	.01	.96

Fig. 6. Confusion matrix for the KTH dataset

We used a leave-one-out procedure specifically we used 24 actors' clips as a training set and the remaining actor's clips as a test set. Performance is presented as the average accuracy of 25 runs, each with a different person. First we tested our descriptor using the nearest-neighbor classifier based on the Euclidean distance and increasing the amount of moments (see Fig. 5). With this approach the use of high order moments degrades the performance of the classifier. This is due to the fact that the high order filters response in small scale cuboids is mostly noisy. Then we used our pyramid similarity kernel increasing the levels of detail. As discussed in Sect. 3 levels with higher order moments are weighted more than levels with lower order moments. We can see that in this case we can exploit the higher details captured by high order moments without degrading the overall classifier performance.

The confusion matrix reported in Fig. 6 shows that as expected jogging and running are the most difficult actions to discriminate while for all other classes results are quite satisfying.

Table 1. Descriptor complexity comparison together with accuracy

Method	Size	Computation time	Accuracy
Pyramid Zernike 3D	84	0.0300 s	91.30%
Gradient + PCA[20]	100	0.0060 s	81.17%
3D SIFT[7]	640	0.8210 s	82.60%
Ext Grad LBP-TOP + PCA[22]	100	0.1000 s	91.25%
3DGrad[9]	432	0.0400 s	90.38%
HOG-HOF[3][5]	162	0.0300 s	91.80%
HOG3D[3][8]	380	0.0020 s	91.40%
SURF3D[3][6]	384	0.0005 s	84.26%

In Tab. 1 we compare our descriptor with the state-of-the-art on KTH dataset respect to the computation time, storage needs and accuracy. Computation time is measured on our machine when the code was available while it is reported from the original publication if not. The accuracy is reported from the experiments reported in the original publication. We can see that Pyramid Zernike 3D descriptors are the smallest in terms of storage and are fast as other non-trivial implementations and C/C++ implementations; note that Gradient PCA is a simple concatenation of pixel gradient values and projection on principal components. Our descriptor is implemented without any optimization in MATLAB.

6 Conclusions

In this paper we have presented a method for action classification based on a new compact descriptor for spatio-temporal interest points. We introduce a new kernel suitable for matching descriptors that can be decomposed in multi-resolution sets. The approach was validated on the KTH dataset, showing results that have a low spatial and temporal computational complexity with comparable performance with the state-of-the-art. Our future work will deal with evaluation on more realistic datasets.

Acknowledgments. This work is partially supported by the EU IST euTV (contract FP7-226248).

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
2. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proc. of ICCV (2003)
3. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A comparison of affine region detectors. International Journal of Computer Vision 65(1-2) (2005)

[3] c++ implementation.

4. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(10) (2005)
5. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proc. of CVPR (2008)
6. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. ECCV, pp. 650–663. Springer, Heidelberg (2008)
7. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proc. of ACM Multimedia (2007)
8. Kläser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: Proc. of BMVC (2008)
9. Ballan, L., Bertini, M., Del Bimbo, A., Seidenari, L., Serra, G.: Recognizing human actions by fusing spatio-temporal appearance and motion descriptors. In: Proc. of ICIP (2009)
10. Flusser, J., Zitova, B., Suk, T.: Moments and Moment Invariants in Pattern Recognition. Wiley Publishing, Chichester (2009)
11. Li, S., Lee, M.C., Pun, C.M.: Complex zernike moments features for shape-based image retrieval. IEEE Transactions on Systems, Man, and Cybernetics (2009)
12. Sun, X., Chen, M., Hauptmann, A.: Action recognition via local descriptors and holistic features. In: Proc. of Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB) (2009)
13. Bo, L., Ren, X., Fox, D.: Kernel descriptors for visual recognition. In: Advances in Neural Information Processing Systems (2010)
14. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. of CVPR (2006)
15. Grauman, K., Darrell, T.: The pyramid match kernel: Discriminative classification with sets of image features. In: Proc. of ICCV (2005)
16. Neri, A., Carli, M., Palma, V., Costantini, L.: Image search based on quadtree zernike decomposition. Journal of Electronic Imaging 19(4) (2010)
17. Li, S., Lee, M.C., Pun, C.M.: Complex zernike moments features for shape-based image retrieval. IEEE Transactions on Systems, Man, and Cybernetics 39(1) (2009)
18. Canterakis, N.: 3d zernike moments and zernike affine invariants for 3d image analysis and recognition. In: Proc. of Conference on Image Analysis (1999)
19. Novotni, M., Klein, R.: Shape retrieval using 3d zernike descriptors. Computer-Aided Design 36(11), 1047–1062 (2004)
20. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: Proc. of VSPETS (2005)
21. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: Proc of. CVPR (2008)
22. 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)

A Low Complexity Motion Segmentation Based on Semantic Representation of Encoded Video Streams

Maurizio Abbate, Ciro D'Elia, and Paola Mariano

Università di Cassino
Via G. Di Biasio, 43 02043 Cassino (FR) – Italy
{m.abbate,delia,p.mariano}@unicas.it

Abstract. Video streaming is characterized by a deep heterogeneity due to the availability of many different video standards such as H.262, H.263, MPEG-4/H.264, H.261 and others. In this situation two approaches to motion segmentation are possible: the first needs to decode each stream before processing it, with a high computational complexity, while the second is based on video processing in the coded domain, with the disadvantage of coupling between implementation and the coded stream. In this paper a motion segmentation based on a "generic encoded video model" is proposed. It aims at building applications in the encoded domain independently by target codec. This can be done by a video stream representation based on a semantic abstraction of the video syntax. This model joins the advantages of the two previous approaches by making it possible working in real time, with low complexity, and with small latency. The effectiveness of the proposed representation is evaluated on a low complexity video segmentation of moving objects.

1 Introduction

In many applications the video stream is available at the source only in compressed form because it is included in some system protocols such as 3G-324M, DVB-H, SIP/RTP and others, or because the NetCams, nowadays used in many applications, send data on RTP/RTSP in MPEG-4, H.264, H.263 and others. In this varied context, a direct approach to implement video processing algorithms consist in a work in the pixel domain after video decoding. This choice has the disadvantage of an high complexity, because of decoding and pixel domain processing. Moreover the encoded stream contains useful information for many applications, such as motion vectors. Hence a processing using such information is able to fully exploit the work of the encoder. For these reasons the proposed approach detects moving regions in the compressed domain without decompression. This allows us to work in real time, with low complexity, and with a small latency because, in principle, the representation could be constructed on the head of a video picture while its tail is still being received. The proposed representation addresses also the problem that in video applications,

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 209–218, 2011.

as stated before, there are many encoded stream standards, thus implying some problems such as the dependence between the implementation and the codec, or in other words the leak of generality of the implemented algorithm. Indeed our proposal is to represent the video using a generic model of encoded video streams, in order to develop codec independent algorithms. In this paper we discuss a low complexity motion segmentation, based on a semantic and unified representation of encoded video streams. Motion segmentation is a process that decomposes a video scene into moving objects or regions. Moving objects or regions extraction is a necessary preprocessing for many application such as scene interpretation, analysis and other. Moreover, the motion segmentation is also worthwhile in cooperation with real-time video trans-coding, object recognition, video surveillance applications and source encoding with content based video standards. Many segmentation methods proposed in the literature, regard the extraction of motion information as a separate process without using any information derived from video encoding process as [9,6]. Other methods use the information derived from encoded bitstream, like DCT coefficients [14,5,10,7], motion vectors [2,12,4,3,8], or both [11,13,1]. These are efficient for real-time applications because they work on encoded domain, however they have the disadvantage of being codec dependent. The novelty of our method is to build a video segmentation, independently from the codec syntax by the use of a semantic model of the encoded stream. The motion segmentation presented in this work is block-based and computed through some basic features contained on the semantic representation, like motion vectors, intra/inter block information and others. The direct use of these features is not sufficient for our aims, indeed encoders operate given a bit budget in order to minimize the source distortion. This implies that some features contained in the bitstream, like motion vectors, are present not only on moving regions but also on similar areas in order to efficiently represent some local changes, using information of neighborhood areas. In other circumstances on flat areas those features are absent even on moving objects. The approach has been evaluated on some standard sequences, but due to space constraints, we present results on two of the most used test videos, "Hall Monitor" and "Container", in order to make easy the comparison in literature. The results show that real-time motion segmentation can be achieved by using our approach. The paper is organized as follows. Section 2 presents the video bitstream model, Section 3 shows details of the proposed motion segmentation method. Experiments, conclusions and future works are in Section 4 and Section 5.

2 Encoded Video Stream Model

The proposed encoded video stream model is based on the abstraction of syntax elements of the most significant video standards. Looking at the standards and at their evolution, it's easy to notice that they share some common ideas. Indeed video standards syntax evolution moves from simple to more and more complex scene models, or on more complex tools for representing it. Video standard syntax could be seen as a set of encoding tools that, if appropriately used,

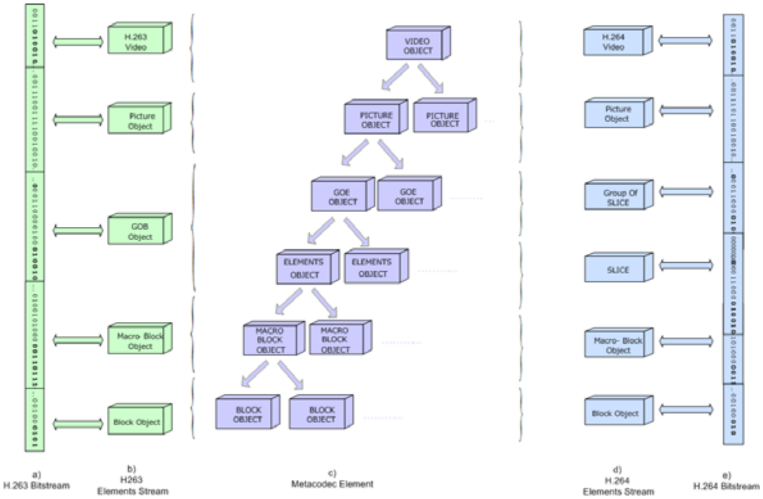


Fig. 1. Assuming in a) and e) respectively an H.263 and a H.264 bitstreams, the semantic representations specific for H.263 and H.264 bitstreams are shown respectively in b) and d), in which the hierarchical dependencies between elements defined by this standards are maintained; in c) the Semantic Representation of Encoded Video Stream is shown, which is able of generalizing different video standards

let the encoder to achieve "good" compression. The aim of our model is to catch common elements of the main standards or to generalize the novelty introduced by some of them. For this reason, introducing the proposed model is useful to reconsider some concepts about video compression and some encoding tools provided by various standards. As it is well known, video itself is basically a three-dimensional array of color pixels of which two dimensions lies in the spatial directions and the last dimension represents the time evolution of the scene. Video data contain spatial, temporal and chromatic redundancy, hence, in order to achieve video compression, standard codecs provide different tools to exploit redundancies. For the spacial redundancies the use of DCT transform tools is quite common. There are many syntactic differences, which could be viewed as a different scene model or different tools to represent the same model. More in details the syntax tools provided in ITU-H.261 are adequate on a scene model with a main object on a background, for instance a speaker on a background. Indeed H.261 introduces encoding of "future frame" by P-frame encoding tool. ITU-H.263 is also suited on the same scene model, but introduces more complex tools for representing scene evolution like PB-frames (Bidirectional Prediction), and many others. The ISO-IEC MPEG-2/ITU-H.262 has encoding tools to predict with high efficiency the redundant information by P-frames and B-frames and also introduces some tools to encode interlaced analog signals. The ISO-IEC MPEG-4 part2 introduces, instead, more complex scene models, made up by several visual object on the scene, explicitly encoded in the bitstream but quite complex to exploit. Finally ISO-IEC MPEG-4 part10/ITU-H.264 uses

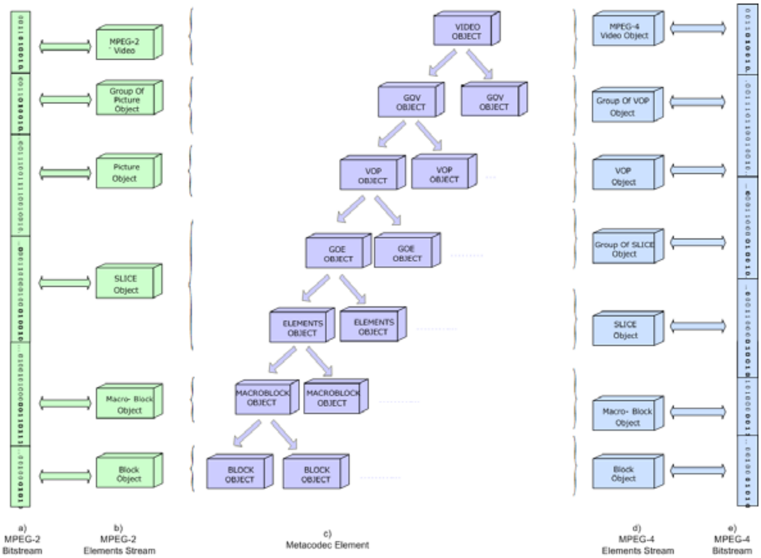


Fig. 2. Assuming in a) and e) respectively an MPEG-2 and a MPEG-4 bitstreams, the semantic representations specific for of MPEG-2and MPEG-4 bitstreams are shown respectively in b) and d), in which the hierarchical dependencies between elements defined by this standards are maintained; in c) the Semantic Representation of Encoded Video Stream is shown, which is able of generalizing different video standards

advanced but easy to implement, tools to represent the evolution of the scene. These tools give it a big advantage if compared with other standards in terms of performance. Even with this short survey of same characteristics of the main codec standards, it is possible to understand that, to obtain a unified semantic representation of the "encoder tools' syntax", is useful to build an abstraction of the different implementations of the encoding tools themselves such as cited P/B-frames. For instance the structure shown in Fig. 1b, represents the semantic description of the H.263 bitstream (Fig. 1a), where the hierarchical dependencies between elements defined by the H.263 standard are maintained. This type of representation makes video more easily manageable, but it's valid only for the H.263 bitstream. Indeed an H.264 bitstream cannot be represented by the same structure. To this aim we should define a specific representation in which groups of slices and slices appear (see Fig. 1d). In Fig. 1c) the generic model proposed to represent an encoded video stream is shown, where the sought level of generalization / abstraction is obtained through each of the semantic elements which compose the model. Each element of the model generalizes one or more syntactical elements of different video standards.

The model consists of Picture Objects containing information about the frame format (CIF, QCIF, etc.), the coding type used (Intra, Inter, etc.) and the quantization step size. *GroupOfElements* Object handles the case where the frame is divided into *SliceGroup* (H.264, MPEG-2). *Elements* Object generalizes GOB

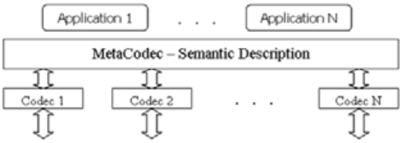


Fig. 3. Framework layers: the low layer is specialized on the supported codecs, the middle layer performs the semantic representation and finally on the top the application layer

(H.261, H.263) and Slices (H.263+, MPEG-2, H.264). The *Macroblock* Object contains the coding type used in the macroblock, the motion vector value and the blocks transmitted (Coded Block Pattern). Finally the *Block* Object contains luminance (Y) and chrominance (Cr-Cb) DCT coefficients. Though the proposed abstraction model is valid only for H.26X encoding, it can be extended to MPEG encoding streams by providing the semantic structure with a cap consisting of elements generalizing the Video Object Planes (VOP) and groups of VOPs, or groups of Pictures, which are typical syntactic elements of MPEG standards. Therefore we propose a model which provides a generic representation of the video bitstream at a higher level of abstraction, i.e. a metacodec (see Fig. 3). The framework is organized in three main layers: the low layer specialized on the supported codecs, which should map the specific codec syntax into the semantic representation of the middle layer, called "Metacodec". On the top, at the third level, there is the application layer, like segmentation, that is based on "Metacodec" layer, which, being codec independent, allows the development of codec independent applications/algorithms. Hence the video segmentation, discussed in details in Section 3, places itself at the third level of Fig. 3. In order to easily evaluate the results of the motion segmentation algorithm, we have implemented a visualizer algorithm, called Halo, based on the semantic representation of the encoded stream, therefore codec independent. Experimental results show the effectiveness of the abstraction model.

3 Motion Segmentation Based on Semantic Description

In the context of video analysis, processing and coding, the segmentation is a very common operation and of particular interest. Video segmentation aims at partitioning a video sequence into several regions according to a given criterion. In the context of video analysis, processing and coding, the segmentation is a very common operation and of particular interest. As said before segmentation is at the third level of Fig. 3, but the segmentation itself could be seen as a further semantic layer, between applications and MetaCodec, in particular as a low level layer for the scene content representation. The presented motion segmentation algorithm starts from the basic idea that motion is a special change. It works in two steps: Spatial Segmentation and Temporal Segmentation. The *Spatial Segmentation* step labels a macroblock (MB) of a frame according to the MB spatial changes, then the *Temporal Segmentation* labels a MB according to its temporal behavior.

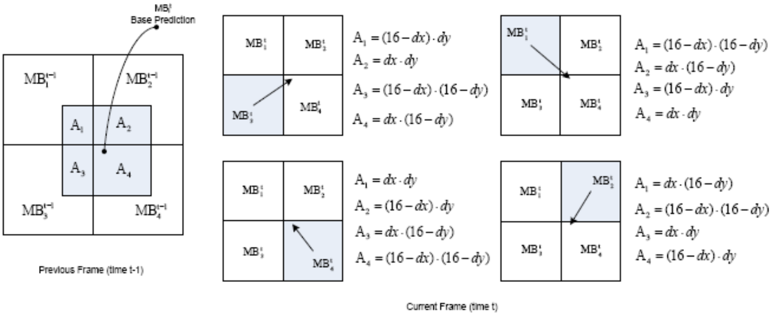


Fig. 4. Temporal evolution of a MB "active", where dx and dy indicate the horizontal and vertical components of MV_i^t (MV_x, MV_y) associated to MB_i^t

3.1 Spatial Segmentation

In the first segmentation step we take into consideration the MBs of each frame. A MB that represents a moving object or local changes in the background is tagged as 1 ("active"), while a MB that is not changing is tagged as 0 ("static"), according to the following equation:

$$ACT_i^t = \begin{cases} 1, \begin{cases} \text{if } MV_i^t \neq 0. \\ \text{if } MB_i^t \text{ is INTRA in INTER frame.} \\ \text{if } MV_i^t = 0 \text{ and neighborhood is} \\ \text{``mainly'' in one of the previous cases.} \end{cases} \\ \\ 0, \ \text{otherwise.} \end{cases} \tag{1}$$

where ACT_i^t, MB_i^t and MV_i^t, are respectively the action tag, the macroblock and the motion vector of the $i-th$ block at time t. More in deep, if the motion vector is zero, we consider the eight adjacent MBs of the current MB; if the neighborhood consists of a number of inter or intra MBs greater than a threshold Th_1, the MB considered is labeled as "active", otherwise "static".

3.2 Temporal Segmentation

After Spatial Segmentation, Temporal Segmentation has been performed in order to understand if an area is "active" due to a moving object or to a local change.Temporal Segmentation operates on matrices of labels ("static" or "active") previously computed. Starting from the consideration that a MB belonging to a moving object has a coherence of the action through some frames, at this stage our aim is to assess the temporal behavior of the "active" INTER MBs. We can say that it is possible to calculate roughly the Motion Coherence (MC) of the current MB at time t using the frames at time $t-1$ and $t-2$ through the weighted average of ACT coefficients influencing the current MB. More in deep,

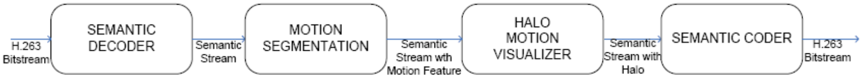


Fig. 5. The diagram blocks sequence of the experimental phase

$MC(1)_i^t$ and $MC(2)_i^t$ are defined as the weighted average of the ACT coefficients influencing MB_i^t block respectively at time $t-1$ and $t-2$, as in the following definitions:

$$MC(1)_i^t = \frac{\sum\limits_{k \in \eta(i,MV_i^t)} ACT_k^{t-1} \cdot A_k(i, MV_i^t)}{\sum_{k=1}^4 A_k(i, MV_i^t)} \tag{2}$$

$$MC(2)_i^t = \frac{\sum\limits_{k \in \eta(i,MV_i^t)} MC(1)_i^{t-1} \cdot A_k(i, MV_i^t)}{\sum_{k=1}^4 A_k(i, MV_i^t)} \tag{3}$$

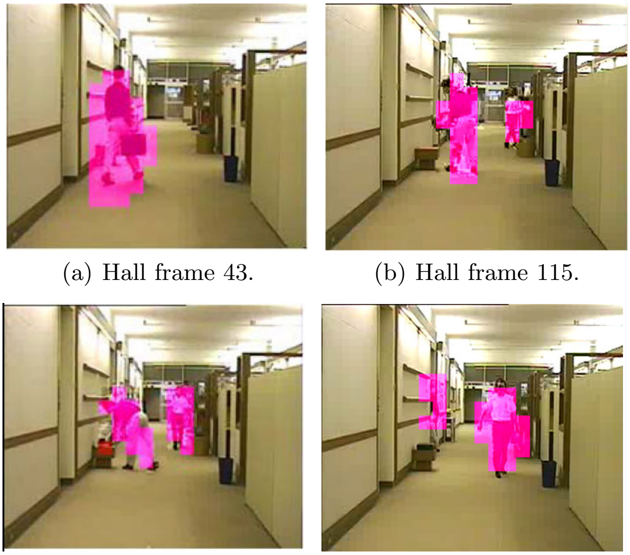
where $\eta(i, MV)$ is the set of indices of the MBs, in the previous frame, influencing the MB of index i by Motion Vector MV, depicted as $\{1, 2, 3, 4\}$ in Fig. 4 for sake of simplicity. In the same figure the current MB_i^t is predicted from parts of the four macroblocks MB_k^{t-1}, furthermore the coefficients $A_k(i, MV)$, representing the influence of each of the four macroblocks MB_k^{t-1} on the current macroblock, are calculated through the motion vector associated to the MB_i^t block, as shown by the formulas, where dx and dy are respectively the horizontal and vertical components of the motion vector. Finally, the temporal segmentation algorithm decides whether the MB_i^t belongs to a moving object depending on the fact that the following equation is greater or smaller than a threshold Th_2:

$$MOV_i^t = \alpha_0 \cdot ACT_i^t + \alpha_1 \cdot MC(1)_i^t + \alpha_2 \cdot MC(2)_i^t \tag{4}$$

where $\alpha_0 + \alpha_1 + \alpha_2 = 1$ and MOV is close to 0 if there is no movement and close to 1 if there is a coherent movement from previous MBs to the current one.

4 Experimental Results

The proposed motion segmentation was tested on some video sequences recorded by a static camera. For briefness, only the results obtained for the “Hall Monitor” (300 frames) and “Container” (300 frames) CIF sequences will be discussed. Each sequence was encoded into a simple H.263 profile. The encoded video bitstream was decoded by the metacodec layer and then the result was used by our algorithm of motion segmentation to detect the moving objects of the scene from a largely static background (Section 3). Because the use of a basic H.263 codec, the resulting segmentation is related with the size of macroblock (16×16 pixel). It was found that, by applying 300 frames “Hall” sequence to the proposed motion segmentation, coherent motions are actually detected throughout the whole



(a) Hall frame 43.

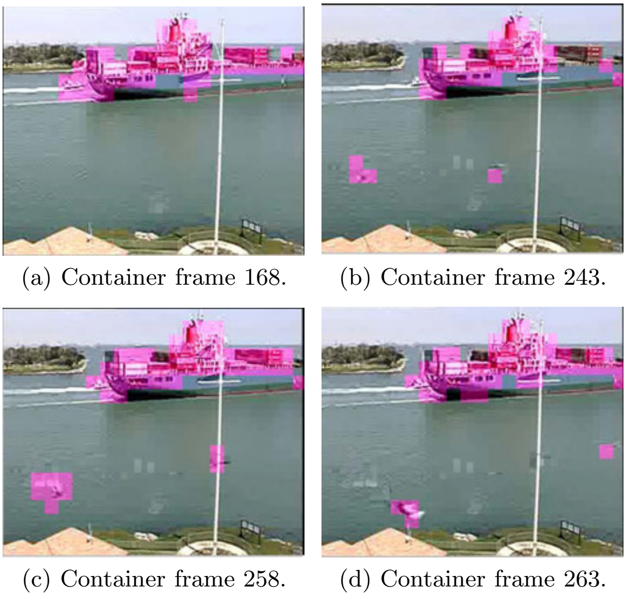
(b) Hall frame 115.

(c) Hall frame 150.

(d) Hall frame 241.

Fig. 6. Experimental Results on "Hall Monitor" Video Sequence ($Th_1 = 3$, $Th_2 = 0.6$, $\alpha_0 = 0.4$, $\alpha_1 = 0.3$, $\alpha_2 = 0.3$)

video sequences. The motion segmentation's results are visualized through the "Halo" algorithm, cited in Section 2, as the block diagram in Figure 5) shows. Figure 6 shows the segmentation results in the 43-th (man in black T-shirt), 115-th (entry of man in white T-shirt), 125-th (both men in the corridor), 241-th (output of man in black T-shirt) frames respectively; the segmentation algorithm detects the real movement of video scene, i.e. the two people moving in the corridor. Figure 7 shows the segmentation results for the "Container" video sequence in the 168-th (ship and speedboat), 243-th (passage of both gulls) and 268-th (exit of the first gulls) frames respectively. In the 168-th frame the segmentation is able to distinguish the two boats, as it is able to distinguish the two gulls in the 243-th frame. By examining all the experiments (not only those discussed here), we have observed that: 1) parameters α_0, α_1 and α_2 respectively set to 0.4, 0.3 and 0.3 are suitable for all tested video sequences; 2) the threshold Th_1 allows us to select the dimension of the moving objects we want to detect; 3) if we set the threshold Th_2 to 0.6, we can apply a majority rule to decide if a macroblock belongs to a moving object (i.e. if it is a moving macroblock in at least two of three frames). The experimental results show that our segmentation is able to identify motion areas using the limited information previously calculated by the codec (motion vectors and macroblock type) with very simple decision rules and a low computational complexity. However, it can wrongly decide like in the 168-th frame of Figure 7, where the algorithm detects motion of waves caused by the transit of the motorboat, so representing in this case the small changes of



(a) Container frame 168.

(b) Container frame 243.

(c) Container frame 258.

(d) Container frame 263.

Fig. 7. Experimental Results on "Container" Video Sequence ($Th_1 = 3$, $Th_2 = 0.6$, $\alpha_0 = 0.4$, $\alpha_1 = 0.3$, $\alpha_2 = 0.3$)

the background instead of the movements related to a real object. This kind of error is not due to a malfunctioning of the segmentation algorithm but simply to the fact that we are using only the information about the movement. Indeed the codec sometimes associates motion vectors to macroblocks that are not moving but are changed in texture; many of the errors due to this reason have been eliminated with the temporal segmentation, by recognizing as "active" only those macroblocks whose motion remains consistent through three frames.

5 Conclusion and Future Developments

In this paper, a semantic model of encoded bitstreams has been proposed. Based on this model, a generic block-based motion segmentation method is introduced. The segmentation results from the Hall Monitor and Container sequences show that the proposed method can exploit the semantic description of an encoded video to perform motion segmentation. It also offers a practical approach to integrate the video encoding with the motion segmentation process, which indicates that the proposed segmentation method is suitable for real-time video applications like video surveillance and video transcoding. Future studies will be focused on detecting the video contents and their features through an efficient video segmentation algorithm based not only on motion but also on texture features computed from the semantic description.

References

1. Ahmad, A., Ahmad, B., Lee, S.: Fast and robust object detection framework in compressed domain. In: Proc. IEEE Sixth Int. Symposium on Multimedia Software Engineering, pp. 210–217 (December 2004)
2. Chung, R., Chin, F., Wong, K., Chow, K., Luo, T., Fung, H.: Efficient block-based motion segmentation method using motion vector consistency. In: MVA 2005 IAPR Conference on Machine Vision Applications, pp. 550–553 (May 2005)
3. Hong, W., Lee, T., Chang, P.: Real-time foreground segmentation for the moving camera based on h.264 video coding information. In: Proc. IEEE Int. Conf. on Future Generation Communication and Networking, pp. 385–390 (December 2007)
4. Hsieh, C., Lai, W., Chiang, A.: A real time spatial/temporal/motion integrated surveillance system in compressed domain. In: Proc. IEEE Int. Conf. on Intelligent Systems Design and Applications, pp. 658–665 (November 2008)
5. Ji, S., Park, H.: Region-based video segmentation using dct coefficients. In: Proc. IEEE Int. Con. Image Processing, vol. 2, pp. 150–154 (October 1999)
6. Karayiannis, Varughese, Tao, Frost, Wise, Mizrahi.: Quantifying motion in video recordings of neonatal seizures by regularized optical flow methods. IEEE Trans. Image Process.14(7), 890–903 (July)
7. Lee, S.W., Kim, Y.M., Choi, S.W.: Fast scene change detection using direct feature extraction from mpeg compressed video. IEEE Trans. Multimedia 2(4), 240–254 (2000)
8. Neri, A., Colonnese, S., Russo, G., Talone, P.: Automatic moving object and background separation. Signal Process.(Special Issue) (66), 219–232 (1998)
9. Nguyen, H., Worring, M., Dev, A.: Detection of moving objects in video using a robust motion similarity measure. IEEE Trans. Image Process. 1(9), 137–141 (2000)
10. Pons, J., Prades-Nebot, J., Albiol, A., Molina, J.: Fast motion detection in compressed domain for video surveillance. IEEE Electronics Letters 38(9), 409–411 (2002)
11. Porikli, F., Bashir, F., Sun, H.: Compressed domain video object segmentation. IEEE Trans. Image Process. 1(5297), 2–14 (2010)
12. Ritch, M., Canagarajah, N.: Motion-based video object tracking in the compressed domain. In: Proc. IEEE Int. Con. Image Processing, vol. 6, pp. 301–306 (2007)
13. Tao, K., Lin, S., Zhang, Y.: Compressed domain motion analysis for video semantic events detection. In: Proc. IEEE Int. Conf. on Information Engineering, pp. 201–204 (July 2009)
14. Zeng, W., Gao, W., Zhao, D.: Automatic moving object extraction in mpeg video. In: Proc. IEEE Int. Symposium on Circuits and Systems, vol. 2, pp. 524–527 (2003)

Audio-Video Analysis of Musical Expressive Intentions

Ingrid Visentini[1], Antonio Rodà[1], Sergio Canazza[2], and Lauro Snidaro[1]

[1] University of Udine, Dept. of Mathematics and Computer Science, via Margreth 3,
33100 Udine, Italy
[2] University of Padova, Dept. of Information Engineering, Via Gradenigo 6/B,
35131 Padova, Italy
{ingrid.visentini,antonio.roda,lauro.snidaro}@uniud.it,
canazza@dei.unipd.it

Abstract. This paper presents a preliminary study on the relation between audio-video streams and high-level information related to expressive nuances. A violinist was asked to play three musical excerpts several times, each one inspired by one of nine different expressive intentions. Perceptual tests was carried out using both audio-only and audio-visual recordings of the performances. The results demonstrate that the visual component aids the subjects to better recognize the different expressive intentions of the musical performances, showing that the fusion of audio-visual information can significantly improve the degree of recognition given by single means.

1 Introduction

This paper presents a preliminary study on the relation between audio-video streams and high-level information related to expressive nuances. At the moment, our concern is focused on audio-visual recordings of musical performances, as they have interesting applications both in multimedia information retrieval and in performing art contexts.

The sharing of increasingly large digital audio-visual libraries of musical performances over the network demands sophisticated tools to enable users to easily find the requested content. The textual approach used by today's search engines has limitations in its application to audio-visual files, because it allows only searching by metadata (i.e., title, author, genre, and so on), not by content. So if metadata, which are usually added manually, are incorrect or do not match with the content, the search can fail. Moreover, the user may not know exactly what document he/she is looking for, but might want to browse the audio-visual library to search for a musical performance that meets certain criteria: for example, a relaxing content or ”something hard”. In recent years, much progress has been made toward developing tools for content-based retrieval in audio-only documents (see [6] and [10] for a review). One of the most used approaches is to define a set of features that describe certain characteristics of sound and can be used to automatically classify the songs according to a determined list of categories. Almost completely unexplored is the joint use of audio and video analysis to improve the classification task.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 219–228, 2011.

Much contemporary music can take advantage of multi-modality to enhance performance as a globally engaging experience: music can be considered a conveyor of expressive content related to performance gestures. Several audio-visual operas (e.g. ”Medea” by Adriano Guarnieri) explicitly insist on wanting to achieve an expressive matching between instrumental gestures and physical movement, so that both gestures would reinforce each other producing a more powerful and complete message. Although our work is at a basic research level, we believe that a deeper understanding of the relations among musical and video stimuli may improve the design of multisensorial interfaces, towards an effective mediation technology for music creation/production and content access/fruition.

In particular, this paper aims to study the gestures in relation of musical performances inspired by different *expressive intentions* [3]. This term refers to the expressive nuances that a musician wants to convey by means of its performance and includes emotions, affects as well as other sensorial aspects of a gesture. The relation between music and emotions has been largely investigated by the scientific community (see [7] and [8] for review). Mion & De Poli [9] asked three musicians to play several times a few short melodies, following different expressive intentions described by a set of affective and sensorial adjectives. A set of features, considered to be particularly representative of the expressive nuances of the performances, were extracted on the base of a frame size of 4 seconds. Results showed that, using the selected features, a linear classifier can recognize the expressive intentions of the songs, with an accuracy better than chance. Not many, however, are the studies that analyse the movements related with the expressive intentions in music. Camurri et al. [2] defined a multi-layer model to represent common characteristics of different sensorial domains, such as sounds and physical gestures. Dahl and Friberg [5] sudied the role of the different body parts in conveying emotional intentions during music performance, finding that head movement plays an important role in the communication of emotional content. Castellano et al. [4] was asked a pianist to play the same excerpt with different emotionally expressive intentions. The body movements captured by a camera positioned above the performer were analyzed via an automated system capable of detecting the temporal profiles of two motion cues: the quantity of motion of the upper body and the velocity of head movements. Results showed that both were sensitive to emotional expression, especially the velocity of head movements.

This paper has two objectives: i) to verify if an audio-video stream allows to recognize the performer’s expressive intentions better than the audio-only stream; if so, ii) to find a set of descriptors of the video stream, to be related with the expressive intentions. We have recorded some musical excerpts and calculated statistics on both human perception (Section 2.3) and video sequences (Section 2.4). Since our target applications are the libraries of audio-visual and live art performances where we general can not have a control of the shooting condition, we chose to record in a poorly controlled environment, in terms of lighting and viewpoint, and without using markers. This choice has obviously affected the definition of the features to be extracted. As far as it regards video, we extract SURF and Lukas-Kanade features to determine movements and speed of the violinist playing the 27 musical excerpts.

2 Perceptual Experiments

We carried out two perceptual experiments to verify how the recognition of expressive intentions change in the following conditions: 1) audio-only musical stimuli are presented; 2) the audio stimuli are associated to a video of the musician playing the musical excerpts. The underlying assumption is that the video component makes it easy to better discriminate the expressive content of the musical excerpts.

2.1 Material

A violinist was asked to play three musical excerpts several times, each one inspired by one of the expressive intentions described by the following adjectives: happy, sad, angry, calm, hard, soft, heavy, light, and normal. The adjectives were chosen among the most widely used in studies of music performance: four refer to the emotional domain and four to the sensorial one. The normal performance, i.e. a performance that lacks a specific expressive intention, was introduced as a term of comparison to better assess the changes induced by the other expressive intentions. The three musical excerpts were chosen to represent different musical genres: a piece belonging to the Western classical repertoire (the incipit of the Violin Sonata Op. 1 No. 12 by G. F. Haendel), a popular melody (*Twinkle Twinkle Little Star*), and a jazz standard (*I Got Rhythm* by G. Gershwin). The performances were captured by one microphone and the audio signal were recorded in monophonic digital form at 24 bits and 48000 Hz.

Moreover, the musical performances was recorded also by two digital cameras observing the performer form different view angles. The two videos, synchronized with the audio track, have been processed to extract features that could capture the movement of the performer. The idea is then to extract the same type of features from each musical performance and for each expressive variation of it. The final goal is to compare the quantities computed for each variation and see if significant differences are observable that could lead to a discrimination of the expressive intentions of the performer.

The videos were acquired in non-constrained conditions of background and illumination. We extracted and analyzed Lukas-Kanade and SURF features from both videos to capture the movements of the performer.

2.2 Method

The procedure follows the one already used by Bigand in [1]. The experiment was conducted using an especially developed software interface. Participants were presented with a visual pattern of 27 loudspeakers, representing the 27 excerpts in a random order. They were required first to listen to all of these excerpts and to focus their attention on the expressive intention of the excerpts. They were then asked to look for excerpts with similar expressive intention and to drag the corresponding icons in order to group these excerpts. They were allowed to listen to the excerpts as many times as they wished, and to regroup as many excerpts as they wished. Both the experiments were performed by a total of 40 participants. Of these, 20 did not have any musical experience and are referred to as non-musicians and 20 have been music students for at least five years are referred to as musicians.

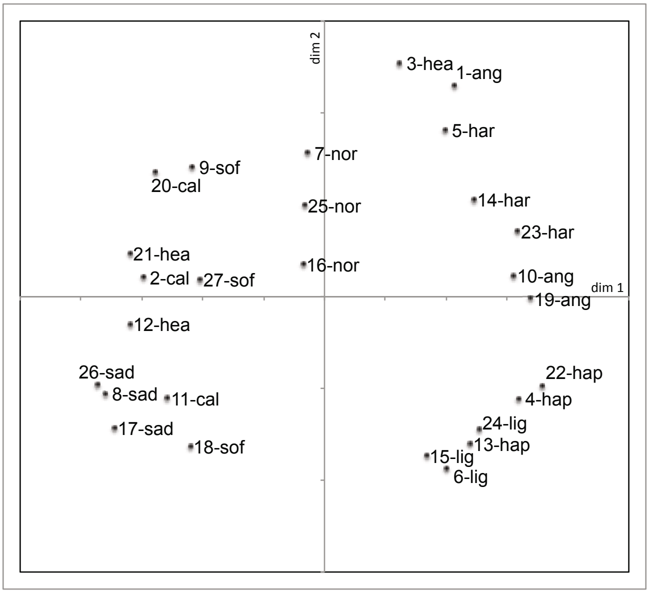


Fig. 1. Multi-Dimensional Scaling of the subjects' answers in the audio-only test

2.3 Results

The excerpts were numerated as follow: from 1 to 9 the performances of the *Violin Sonata* in the order angry, calm, happy, hard, heavy, light, normal, sad, and soft; from 10 to 18 *Twinkle Twinkle Little Star* with the same order; form 19 to 27 *I Got Rhythm*. Participants have formed an arbitrary number N of groups, named G_k. Each group contains the stimuli that the a subject thinks are characterized by the same or a similar expressive intention. The dissimilarity matrix A is defined by counting how many times two excerpts i and j are not included in the same group.

Table 1. Average distance measured between expressive intentions in the audio-only test

	ang	cal	hea	hap	har	lig	nor	sad	sof
ang	28.3	38.7	35.2	29.0	**27.8**	33.2	35.4	38.9	38.4
cal	38.7	**26.7**	32.1	38.7	37.9	37.7	33.7	29.6	27.4
hea	35.2	32.1	33.0	37.9	35.1	37.0	33.7	**30.1**	33.8
hap	29.0	38.7	37.9	**23.0**	32.1	28.2	35.9	39.0	37.9
har	27.8	37.9	35.1	32.1	**27.3**	32.9	33.8	39.1	36.7
lig	33.2	37.7	37.0	28.2	32.9	**26.3**	33.7	38.8	36.8
nor	35.4	33.7	33.7	35.9	33.8	33.7	**27.7**	36.8	31.6
sad	38.9	29.6	30.1	39.0	39.1	38.8	36.8	**21.7**	30.3
sof	38.4	**27.4**	33.8	37.9	36.7	36.8	31.6	30.3	30.3

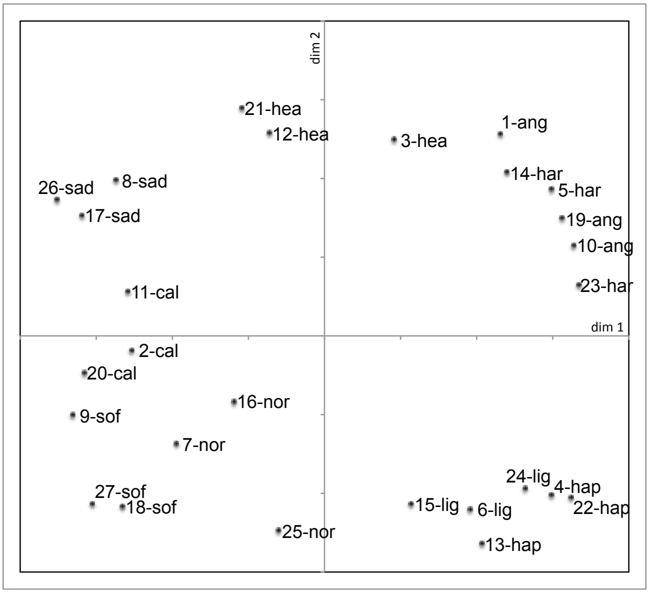


Fig. 2. Multi-Dimensional Scaling of the subjects' answers in the audio-video test

Table 2. Chi-square test on the audio-only test

	ang	cal	hea	hap	har	lig	nor	sad	sof
p-value	0.057	0.023	0.370	5.2e-04	0.038	0.011	0.054	4.4e-05	0.268

I.e., $\forall i, j = 1, ..., 27$ and $\forall k = 1, ..., N$

$$A[i,j] = \begin{cases} A[i,j] + 1 \text{ if} i \in G_k \wedge j \notin G_k \\ A[i,j] \qquad \text{otherwise} \end{cases} \tag{1}$$

Experiment 1: Audio-only Stimuli. The dissimilarity matrix from the experiment 1 was analysed by means of a Multi-Dimentional Scaling (MDS) method. The location of the 27 excerpts along the two principal dimensions is represented in Figure 1. The excerpts that are close in this space are those evaluated to be more similar (in terms of expressive characteristics) by the subjects. It can be noted the three normal performances located in the middle of the two-dimensional space, a cluster composed by the hard and angry performances in the upper right quadrant, a cluster with the happy and light performances in the lower right quadrant, and the three sad performances in the lower left quadrant. On the contrary, heavy and soft performances are not clustered, meaning the subjects did not recognize correctly these expressive intention.

Table 1 shows the average values calculated by grouping the entries of the dissimilarity matrix with the same expressive intention. The performances calm have, among

Table 3. Average distance measured between expressive intentions in the audio-video test

	ang	cal	hea	hap	har	lig	nor	sad	sof
ang	**19.7**	38.7	31.9	33.8	25.3	34.1	37.4	38.8	38.8
cal	38.7	**21.7**	33.0	39.0	38.7	34.7	31.2	31.7	31.3
hea	31.9	33.0	**24.3**	37.7	32.7	38.1	34.7	33.4	35.0
hap	33.8	39.0	37.7	**20.3**	34.1	25.0	35.4	37.6	38.1
har	25.3	38.7	32.7	34.1	**23.7**	33.6	36.3	38.8	38.3
lig	34.1	34.7	38.1	25.0	33.6	**22.3**	33.6	36.9	36.6
nor	37.4	31.2	34.7	35.4	36.3	33.6	**29.3**	34.2	30.8
sad	38.8	31.7	33.4	37.6	38.8	36.9	34.2	**18.7**	33.0
sof	38.8	31.3	35.0	38.1	38.3	36.6	30.8	33.0	**19.7**

Table 4. Chi-square test on the audio-video test

	ang	cal	hea	hap	har	lig	nor	sad	sof
p-value	6.7e-06	9.0e-05	1.9e-03	9.2e-06	8.3e-04	7.7e-04	0.122	3.1e-07	2.7e-06

them, a average dissimilarity of 26.7, which is smaller than the average dissimilarities between calm and the other expressive intentions. The same is true for the expressive intentions happy, hard, light, and sad. In all these cases, a Chi-square test (see Table 2) showed that these values are statistically significant ($p < 0.05$). The subjects' responses regarding the expressive intentions angry, heavy, normal, and soft, instead, are not significant.

Experiment 2: Audio-Video Stimuli. While the results of the experiment 1 show that some expressive intentions are confused, Figure 2 and Tables 3 and 4 show that all the expressive intentions are properly discriminated in the experiment with audio-visual stimuli.

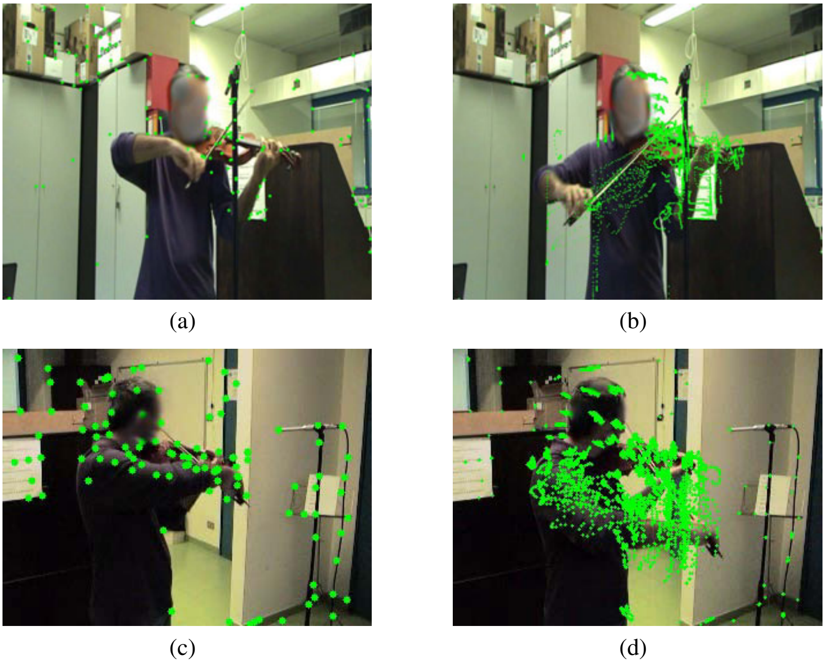
2.4 Video Analysis

An example of feature extraction in video sequences is presented in 3, where the performer movements are followed by the Lukas-Kanade features in the first and second view (Figure 3 (a) and (b) respectively), and the features motion accumulator after 100 frames (in Figure 3 (c) and (d)).

We have dumped the values of the position of each feature for each frame and computed the following statistics:

$$MeanSpread = \frac{1}{n}\sum_{i=1}^{n} \max\left(d_i(x,y)\right) - \min\left(d_i(x,y)\right) \tag{2}$$

$$MeanDerivative = \frac{1}{nT}\sum_{t=1}^{T}\sum\nolimits_{i=1}^{n} |d'_{i,t}(x,y)| \tag{3}$$



(a)

(b)

(c)

(d)

Fig. 3. Examples of frames taken from video sequence, picturing the performer dotted with the Lukas-Kanade features in the first view (a), the features motion accumulator at frame 100 (b), and their behaviour in the second view (c) with the accumulation of their motion after 100 frames in (d)

The first one (2) indicates the average distance covered by the features, calculated as the difference from the maximum and minimum position magnitude. The average is computed on all the features, while the magnitude of the i-th feature is given by

$$d_i(x, y) = \sqrt{x^2 + y^2} \tag{4}$$

where (x, y) is the position of the feature in the image coordinates system. The mean derivative in (3) indicates the average speed of the features. Notice that many features can be stationary since they may be positioned in background regions. For this reason we have not included in the statistics above the features whose "movement" is below a pre-set threshold of 10. In the equations 2 and 3 index i runs from 1 to n over the features in a single frame, while t, where $1 \leq t \leq T$, runs over the total number of frames of the video.

We repeated the test for each musical excerpt, and we obtained the results illustrated in Figure 4 and 5 for first and second camera respectively. In both figures, on the x axis there are the expressive intentions, while on the y axis the mean spread (top graph) and mean derivative (bottom graph) of the features extracted from the video sequence. The standard deviation of the two measures is depicted with an error bar.

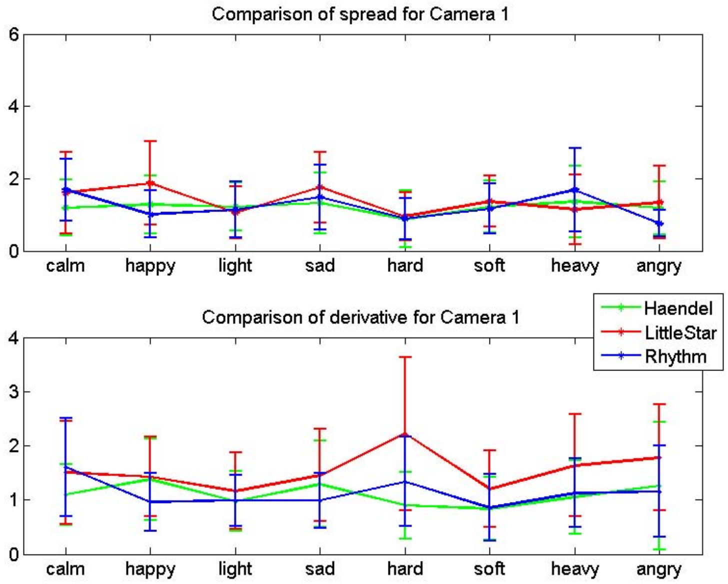


Fig. 4. Mean spread (top) and mean derivative (bottom) of the features extracted from the first camera

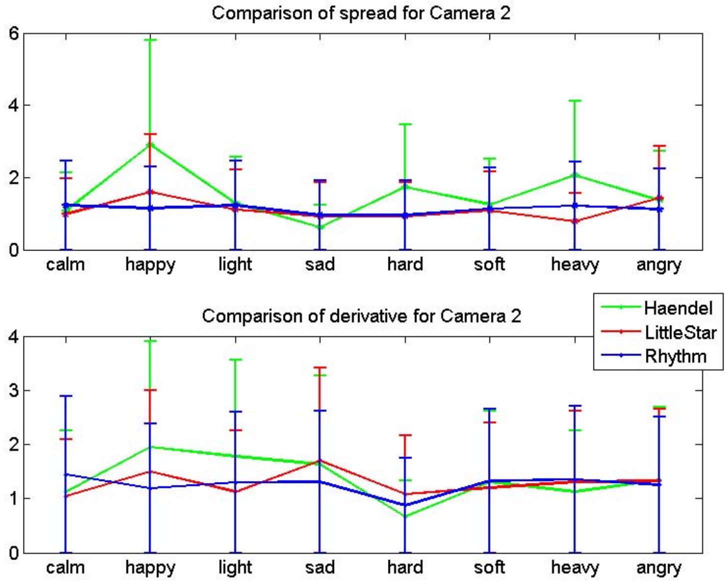
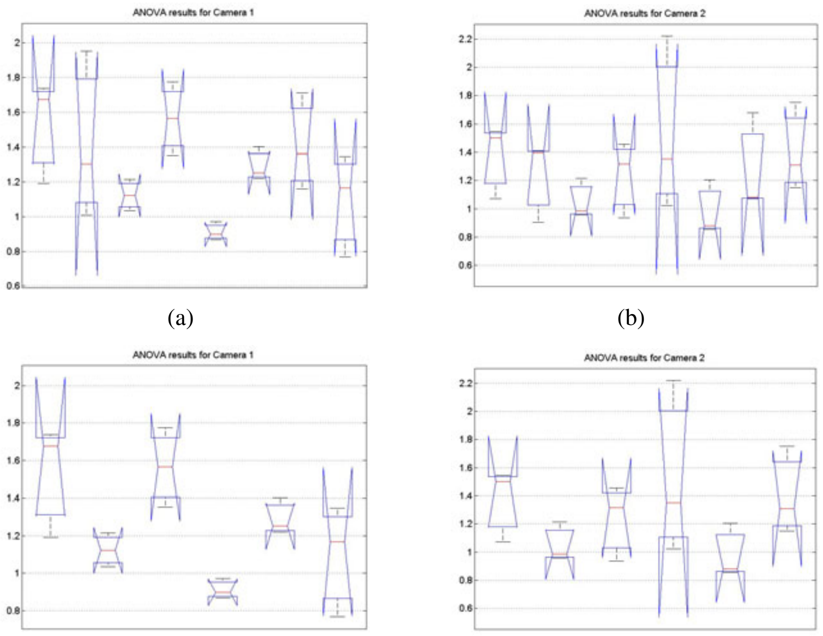


Fig. 5. Mean spread (top) and mean derivative (bottom) of the features extracted from the second camera



(a)

(b)

(c)

(d)

Fig. 6. Results of the ANOVA test for first (a) and second (b) camera on the full expression intentions set. In the second row, the outcomes with a reduced set are presented for the two cameras respectively

We used a one-way ANOVA to test if the expressive intentions are different and separable, considering as null hypothesis the equality of their means. Considering the full set of intentions, we obtained $F(7, 16) = 2.32$ and $p = 0.0773$ for the first camera and $F(7, 16) = 0.9$ and $p = 0.5330$ for the second. Comparing these values with the F-ratio table references, the hypothesis has high probability of being accepted, that is the intentions are not separable. Discarding two expressive intentions with the larger standard deviation, that are *happy* and *heavy* for both cameras, the new values were $F(5, 12) = 4.99$ and $p = 0.0106$ for the frontal view and $F(5, 12) = 1.22$ and $p = 0.3599$ for the other. It is clear that in the case of frontal camera, the removal of the classes with large standard deviation allows to separate the means of remaining expression intentions with a high probability, but in the case of a side camera the situation does not gain much benefit from discarding some observations. Another consideration is that *heavy* and *happy* turn out to be ambiguous expression after automatic video analysis. In Figure 6 are presented illustrations after ANOVA test. In the left column the results for the first camera with all the intentions and with the above mentioned subset are presented. In the second column, the results for the second camera are shown, and the classes are not separable even after the exclusion of the two most uncertain classes.

3 Conclusions

In this paper we presented an analysis of human perception applied to musical expressive intention recognition. The results demonstrate that the visual component aids the subjects to better recognize the different expressive intentions of the musical performances. As humans were tested with audio only, and then with the fusion of audio and video, we selected video features to obtain a fair comparison. We extracted SURF and Lukas-Kanade features to determine movements and speed of the violinist playing 27 musical excerpts. To summarize, the fusion of audio-visual information can significantly improve the degree of expression intention recognition given by single means. Future research direction will be oriented to fuse audio and video features automatically extracted from sequences and to compare them with the results of human recognition.

References

1. Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., Dacquet, A.: Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. Cognition and Emotion 19(8), 1113–1139 (2005)
2. Camurri, A., De Poli, G., Leman, M., Volpe, G.: Communicating expressiveness and affect in multimodal interactive systems. IEEE Multimedia 12(1), 43–53 (2005)
3. Canazza, S., De Poli, G., Rodà, A.: Analysis of expressive intentions in piano performance. Journal of ITC Sangeet Research Academy 16, 23–62 (2002)
4. Castellano, G., Mortillaro, M., Camurri, A., Volpe, G., Scherer, K.: Automated analysis of body movement in emotionally expressive piano performances. Music Perception 26(2), 103–119 (2008)
5. Dahl, S., Friberg, A.: Visual perception of expressiveness in musician's body movements. Music Perception 24, 433–454 (2007)
6. Downie, J.S.: Music information retrieval. Annual Review of Information Science and Technology 37, 295–340 (2003)
7. Juslin, P.N., Sloboda, J.A.: Music and emotion. Theory and research. Oxford University Press, Oxford (2001)
8. Kirke, A., Miranda, E.R.: A survey of computer systems for expressive music performance. ACM Computing Surveys 42(1) (2009)
9. Mion, L., De Poli, G.: Score-independent audio features for description of music expression. IEEE Trans. Speech, Audio, and Language Process. 16(2), 458–466 (2008)
10. Orio, N.: Music retrieval: A tutorial and review. Foundations and Trends in Information Retrieval 1(1), 1–90 (2006)

Image Segmentation Using Normalized Cuts and Efficient Graph-Based Segmentation

Narjes Doggaz and Imene Ferjani

URPAH, Computer Science Department, Faculty of Sciences of Tunis, Tunisia
narjes.doggaz@fst.rnu.tn,
imene.fer@gmail.com

Abstract. In this paper we propose an hybrid segmentation algorithm which incorporates the advantages of the efficient graph based segmentation and normalized cuts partitioning algorithm. The proposed method requires low computational complexity and is therefore suitable for real-time image segmentation processing. Moreover, it provides effective and robust segmentation. For that, our method consists first, at segmenting the input image by the "Efficient Graph-Based" segmentation. The segmented regions are then represented by a graph structure. As a final step, the normalized cuts partitioning algorithm is applied to the resulting graph in order to remove non-significant regions. In the proposed method, the main computational cost is the efficient graph based segmentation cost since the computational cost of partitioning regions using the Ncut method is negligibly small. The efficiency of the proposed method is demonstrated through a large number of experiments using different natural scene images.

Keywords: Image Segmentation, Normalized Cuts, Efficient graph-based, Region adjacency graph.

1 Introduction

Segmentation is defined as the process of partitioning an image into disjoint and homogeneous regions. It is an important step in image analysis. Besides, it is one of the most difficult tasks in image processing, and determines the quality of the final results of the computer vision applications.In the literature, several segmentation methods are defined. They can be classified into three major categories,i.e.,feature based techniques, physics based techniques and image-based techniques. Feature-based methods, such as clustering [4], intend to classify pixels to different groups in a pre-defined color space. These methods gives good results in some cases [19] but their drawbacks are neglecting the spatial information of pixels and the difficulty met in adjusting the number of classes to image regions. The objective of physics based segmentation is to divide an image of a scene in regions that are meaningful in terms of the objects constituting that scene by forming an interpretation of the image based on illumination, material optics and geometry [15]. These methods alleged robustness to highlights,

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 229–240, 2011.

shades and shadowing, but they are not applicable for real images. Image-based techniques are also referred to as region-based when they are based on region entities. Region growing methods [20] are extensively used for this purpose. In such methods regions are iteratively grown by comparing all unallocated neighboring pixels to the regions. These methods can produce an oversegmentation which can be avoided by applying some merging algorithms. Also, graph theoretical methods are considered as region based methods and they have proved their robustness in many image analysis applications, this is why they will be detailed separately.

This paper is organized as follows: in section 2 we explicate the principle of graph theoretical segmentation methods and we focus on efficient graph-based segmentation and normalized cut partitioning algorithm. In section 3, we present and detail our approach that combines the two above methods. In section 4, we present and discuss the results of our approach as well as its computational complexity and we conclude in section 5.

2 Graph Theoretical Segmentation Methods

Graph-based approaches can be regarded as image perceptual grouping and organization methods based on the fusion of the feature and spatial information. The common theme underlying these approaches is the formation of a weighted graph, where each vertex corresponds to an image pixel or a region. The weight of each edge connecting two pixels or two regions represents the likelihood that they belong to the same segment. A graph is partitioned into multiple components that minimize some cost function of the vertices in the components and/or the boundaries between those components. Several graph theoretical-based methods have been developed for image segmentation [24],[23]. Some are region growing methods that join components as a function of the attributes of nodes and edges and others are splitting methods that partition a graph by removing superfluous edges. The efficient graph based method proposed by Pedro et al [10] can be classified as a region growing method. It is highly efficient and cost effective way to perform image segmentation. In splitting methods, the normalized cuts algorithm [17] is extensively used in image segmentation due to its efficiency according to traditional minimum cuts algorithms. However, segmentation based on normalized cuts needs high computational time. Several solutions were presented to solve this problem. Some approaches propose to apply the Ncuts partitioning algorithm to the graph representing a segmented image. By this way there is a great reduction of the computational time since the number of regions obtained by the first segmentation is much smaller than the number of image pixels. But the quality of final segmentation differs from one solution to another. In the work of Makrogiannis et al [13] the image is segmented by applying the watershed algorithm. A graph structure is then applied to represent the relationship between these regions, this resulting graph will be partitioned using Ncuts algorithm. In this method the oversegmentation produced by the watersheds causes a degradation in the region grouping algorithm. To solve this problem Tao et al [18] proposed a similar method which uses mean-shift instead of watershed segmentation, and to avoid inappropriate partitioning, they represent a region with multiple child nodes. In both approaches there is a great reduction of the Ncut cost but the segmentation obtained can not be integrated in real time applications without the need to use a parallel environment. Also, performance in terms of extracting objects from the image is still insufficient. In our work, we use two graph theoretical based methods which are efficient graph based segmentation [10] and the Ncuts partitioning algorithm [17] to provide robust real time segmentation.

2.1 Normalized Cuts Segmentation

A graph-partitioning method attempts to organize nodes into groups such that the intra-group similarity is high and the inter-group similarity is low. Given a graph G = (V,E,W), where V is the set of nodes, and E is the set of edges connecting the nodes, a pair of nodes u and v is connected by an edge and is weighted by $w(u, v) = w(v, u) \geq 0$. $w(u, v)$ measures the dissimilarity between u and v. W is an edge affinity matrix with $w(u, v)$ as its $(u, v)^{th}$ element. The graph can be partitioned into two disjoint sets A and B = V - A by removing the edges connecting the two parts. The degree of dissimilarity between the two sets can be computed as a total weight of the removed edges. In graph theoretic language, it is called the cut[17].

$$cut(A, B) = \sum_{u \in A et v \in B} w(u, v) \tag{1}$$

The problem of finding the minimum cut has been well studied. However, the minimum cut criterion favors grouping small sets of isolated nodes in the graph because the cut defined in equation 1 does not contain any intra-group information. In other words, the minimum cut usually yields overclustered results when it is recursively applied. To avoid this unnatural bias of partitioning out small sets of points, Shi and Malik propose a new measure of dissociation between two groups called Normalized cut(Ncut) [17].

$$Ncut(A, B) = \frac{cut(A, B)}{assoc(A, V)} + \frac{cut(A, B)}{assoc(B, V)} \tag{2}$$

where $assoc(A, V) = \sum_{u \in A, t \in V} w(u, t)$ denotes the total connection from nodes in A to all nodes in the graph, and $assoc(B, V)$ is similarly defined. Unlike the cut criterion that has a bias in favor of cutting small sets of nodes, the Ncut criterion is unbiased [17].

Algorithm 1: ”Normalized cuts” Segmentation[17]

```
Data: I:image, K:number of classes
Result: S:Segmennted image
From I built G=(V,E,W);
Solve $(D-W)y = \lambda Dy$ for eigenvectors with the K smallest
  eigenvalues($(V_1, V_2, ..., V_K)$ {D is a diagonal matrix};
Use the eigenvectors to partition the graph
```

2.2 Efficient Graph-Based Segmentation

The graph based image segmentation is based on selecting edges from a graph, where each pixel corresponds to a node in the graph[10]. Weight on each edge measures the dissimilarity between pixels. The segmentation algorithm defines the boundaries between regions by comparing two quantities:Intensity differences across the boundary and Intensity difference between neighboring pixels within each region. This is useful knowing that the intensity differences across the boundary are important if they are large relative to the intensity differences inside at least one of the regions. This results in a method that obeys certain non-obvious global properties. Let the internal difference of a component C in an image be [10]:

$$Int(C) = \max_{e \in MST(C,E)} w(e) \tag{3}$$

Where $w(e)$ is the largest weight in the Minimum Spanning Tree of the component. Let the difference between two components C_1 and C_2 be the minimum weight edge connecting the two components. That is [10],

$$Dif(C_1, C_2) = \min_{v_i \in C_1, v_j \in C_2, (v_i, v_j) \in E} w((v_i, v_j)) \tag{4}$$

The boundary between a pair of components is determined by checking if the difference between the components, $Dif(C_1, C_2)$, is large relative to the internal difference within at least one of the components, $Int(C_1)$ and $Int(C_2)$. A threshold function is used to determine the degree to which the difference between components must be larger than minimum internal difference.Let the predicate D be [10]:

$$D(C_1, C_2) = \begin{cases} true, & \text{if } Dif(C_1, C_2) > MInt(C_1, C_2); \\ false, & \text{otherwise.} \end{cases} \tag{5}$$

Where the minimum internal difference $MInt$ is defined as [10]:

$$MInt(C_1, C_2) = \min(Int(C_1) + \tau(C_1), Int(C_2) + \tau(C_2)) \tag{6}$$

with

$$\tau(C) = K/|C|$$

$|C|$ is the size of the component C, and K is a constant parameter.

Algorithm 2: "Efficient Graph-based" Segmentation[10]

```
Data: I:image, K:constant
Result: S:Segmented image
From I built G=(V,E), $n$ vertices and $m$ edges;
Sort E into $\pi = (o_1, ..., o_m)$, by non-decreasing edge weight;
$S^0 \leftarrow V$, {each vertex $v_i$ is in its own component};
for $q = 1$ to $m$ do
    if $C_i^{q-1} \neq C_j^{q-1}$ and $w((v_i, v_j)) <= MInt(C_i^{q-1}, C_j^{q-1})$ then
        $S^q$ is obtained from $S^{q-1}$ by merging $C_i^{q-1}$ et $C_j^{q-1}$
    else
        $S^q = S^{q-1}$
    end
end
/*$C_i^{q-1}$ is the component of $S^{q-1}$ containing $v_i$ and $C_j^{q-1}$ the component
  containing $v_j$, $o_q = (v_i, v_j)$                                              */
return [$S = S^m$]
```

3 Our Approach

The outline of our approach can be defined as following. First, an image is segmented into separated regions using the efficient graph based algorithm. Second, the graph representation of these regions is constructed, and the dissimilarity measure between the regions is defined. Finally, a graph-partitioning algorithm based on the Ncuts is employed to form the final segmentation map (fig. 1).

The most important part in our approach is the construction of the region adjacency graph (RAG) and the calculation of the dissimilarity measure between the regions.

The regions obtained by the "Graph-Based" segmentation can be represented by a weighted region adjacency graph G=(V,E,W). The weights on the edges of RAG based on the similarity between two regions play a decisive role in determining the overall performance of our image segmentation process. To define the measure of dissimilarity between neighboring regions, we propose the use of the mean intensity of regions and their compactness [2]. The compactness between two regions is defined as the degree of their adjacency(how much they exhibit adjacency of their constituent parts) since they are considered adjacent when they have at least one adjacent pixel. In our approach, we used the compactness criteria introduced by Adamek which is defined by[2]:

$$C_{ij} = 1 - \frac{F_{i,j}}{\min\{L_i, L_j\}} \tag{7}$$

Where L_i and L_j are the perimeters of regions i and j and $F_{i,j}$ is the length of their common border. C_{ij} gives a value between 0 and 1. When $C_{ij} = 0$ this means that one of regions (i or j) is totally included in the other region. When $C_{ij} = 1$ this means that the regions i and j are not adjacent.

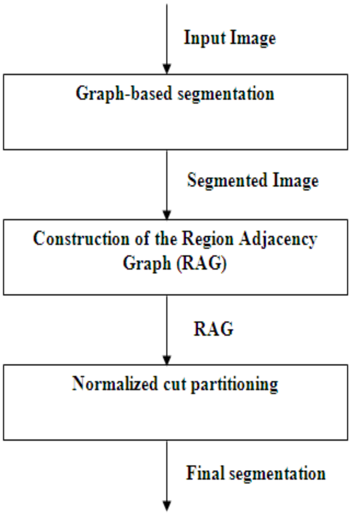

Input Image
Graph-based segmentation
Segmented Image
Construction of the Region Adjacency Graph (RAG)
RAG
Normalized cut partitioning
Final segmentation

Fig. 1. The outline of our approach

We can now propose our definition of the weight $w(i,j)$ between regions i and j by:

$$w(i,j) = \exp(\frac{-\|I_i - I_j\|^2}{\sigma_I^2}) * \begin{cases} \exp - \|C_{ij}\|, & \text{if i et j are adjacents;} \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Where I_i is the mean intensity of region i and the constant σ_I is the same used in the Ncuts segmentation algorithm. The first factor of the above equation is the mean intensity difference between two adjacent regions i and j. The second factor is used to favorise the creation of compact and simple regions.

Once, the region adjacency graph constructed, we apply the normalized cuts algorithm to partition the graph and to group similar regions. The advantages of using hybrid segmentation instead of using separately efficient graph based and normalized cuts segmentation are twofold:

1) It improves segmentation performance. Graph based image-segmentation is a fast and efficient method of generating a set of segments from an image. It not only considers local pixel-based features, but also looks at global similarities within the image but its drawback is that the number of regions generated is relatively high.
2)We can use the Ncuts algorithm as a post processing step without increasing the computational time of the segmentation process because partitioning regions instead of pixels will offer considerable reduction of computational complexity, since the number of image regions is much smaller than that of the pixels. Thus, the size of the weight matrix and, subsequently, the complexity of the graph structure employed for image representation are significantly reduced. Moreover,

applying Ncuts partitioning to regions instead of pixels is more accurate and easy since the edge's weight between tow regions is larger than that between two pixels.

4 Experimental Results

We have applied the proposed algorithm on the Berkeley Segmentation Dataset [3] where the images are 321x481 pixels. We present in fig. 3, fig. 4 and fig. 5 the results of some of our experiments. For each image, the result of the "Graph-Based" algorithm [1], the Ncuts algorithm implemented by Cour et al [7], our segmentation approach and the human segmentation are given. Hence, we can compare visually and intuitively the result of each method of segmentation with the ground truth segmentation.

In the implementation of the "Graph-Based" segmentation there is a pretreatment step to reduce noise. This step needs a smoothness factor σ. Also, there is a post-treatment step that eliminates small regions after applying the segmentation algorithm. The parameter $minsize$ is introduced to define the smallest size that a region can have. The parameters of "Graph-Based" segmentation algorithm are $(K, \sigma, minsize) = (1000, 1, 100)$. We remark that the "Efficient graph-based" segmentation extracts the most important objects in the image but it produces a large number of regions(fig. 3c, fig. 4c and fig. 5c). We remark clearly that the number of these regions in our approach has been decreased in a way that preserves the global structure of the image segmentation. To more illustrate our results let focus in the image of fig. 2 which contains two flowers.

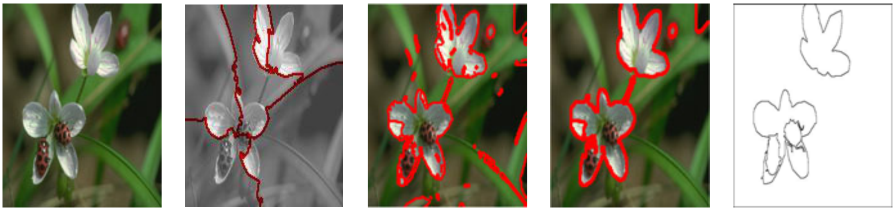


Fig. 2. (a)The original test image. (b)Normalized cuts segmentation. (c)Efficient graph based segmentation. (d)Our proposed method. (e)Human segmentation.

In this example, the segmentation produced by efficient graph based method contains 43 regions. We remark, when we apply our algorithm with setting the number of classes to 6, we obtain a segmentation that preserve the most important objects in the image wich are the two flowers. The regions that belong to the background are also formed into a single region. However, when the Ncut method is directly applied to the image pixels with the same number of classes, the image is partitioned into six regions where each region includes one part of each flower. Our segmentation method distinguishes more objects from the background. By comparing our segmentation results to the ground truth segmentations it is clear that our results are the most similar to the human segmentation. Regions extracted by our method are more meaningful in most of images.

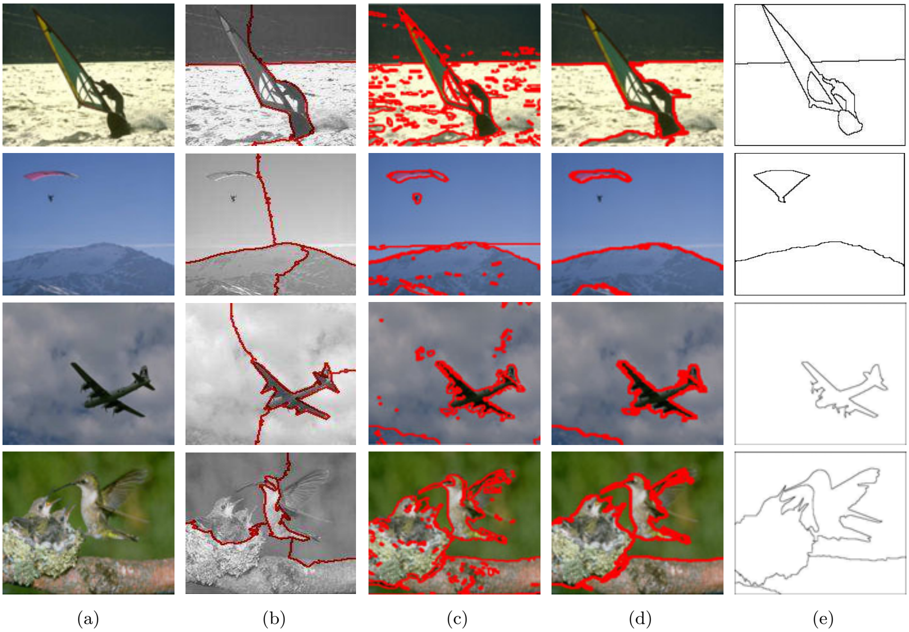


Fig. 3. Test results of different images with partitioning class C=4. (a)The original test images. (b)Normalized cuts segmentation. (c)Efficient graph based segmentation. (d)Our proposed method. (e)Human segmentation.

We evaluate, also, and compare the computational cost of our method with those of the Graph-Based and Ncut ones. For that, we use a PC equipped with a 1.6-Ghz and 1-GB memory. We consider only the first four images from up to bottom in each of fig. 3, fig. 4 and fig. 5. We notice that the computational cost of our approach is near to the "Graph-Based" method which takes 700-800 ms to process an image. This means that the cost of partitioning the region nodes using the Ncut method is negligibly small. However, the normalized cuts segmentation takes 9-19s to segment an image [1].

To evaluate objectively the segmentation results given by our approach we use four different measures that are:

Probabilistic Rand Index (PRI) [21]: PRI counts the number of pixel pairs whose labels are consistent between the segmentation and the ground truth.

Variation of Information (VoI) [16]: VoI measures the amount of randomness in one segmentation that cannot be explained by the other.

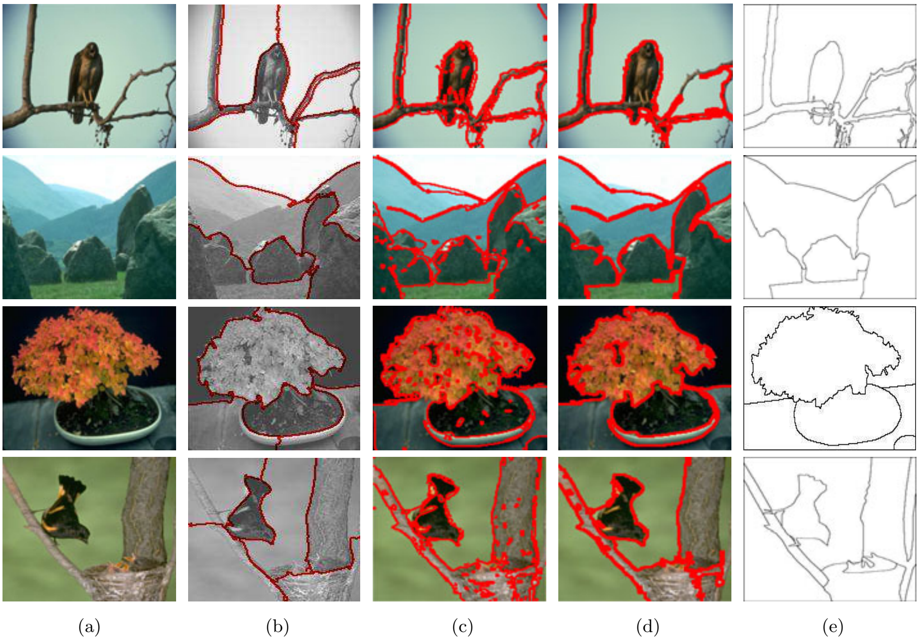

(a) (b) (c) (d) (e)

Fig. 4. Test results of different images with partitioning class C=6. (a)The original test images. (b)Normalized cuts segmentation. (c)Efficient graph based segmentation. (d)Our proposed method. (e)Human segmentation.

Table 1. Comparaison of the computational cost between ”Graph-Based”, Ncuts and Our method

		”Graph-Based” time (s)	Ncuts time(s)	Our method time(s)
C=4	1	0.777	13.297	0.940
	2	0.783	10.023	0.8921
	3	0.782	9.917	0.876
	4	0.779	15.057	0.919
C=6	1	0.775	14.584	0.898
	2	0.814	14.860	0.945
	3	0.764	14.747	0.901
	4	0.832	15.057	0.984
C=8	1	0.770	15.912	0.948
	2	0.784	19.563	0.957
	3	0.821	14.747	0.986
	4	0.765	13.082	0.921

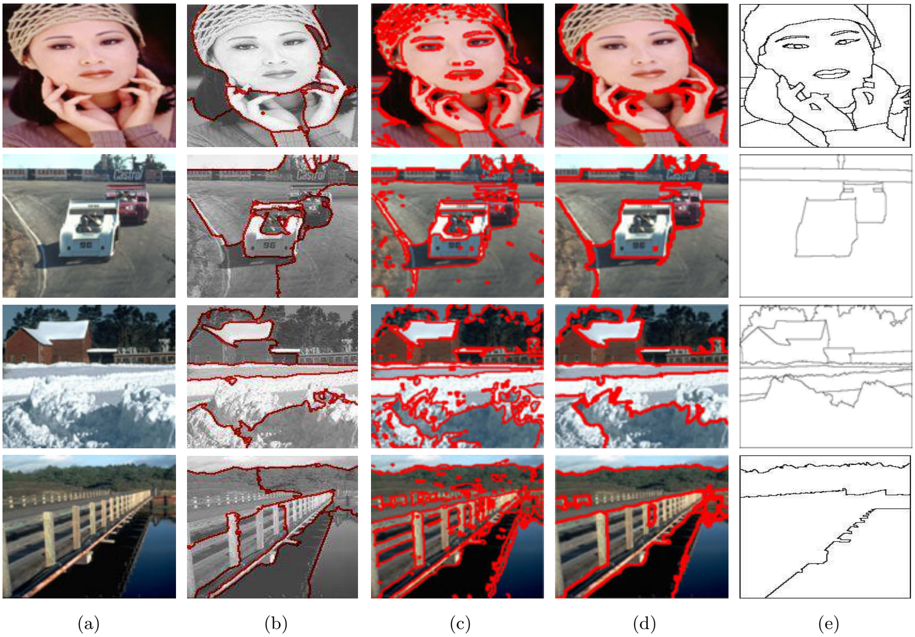


Fig. 5. Test results of different images with partitioning class C=8. (a)The original test images. (b)Normalized cuts segmentation. (c)Efficient graph based segmentation. (d)Our proposed method. (e)Human segmentation.

Table 2. Quantitative comparison of our algorithm with other segmentation methods over the Berkeley database. The best three results are highlighted in colors: Red, Green, and Blue in descending order.

Method \ Score	PRI	VOI	GCE	BDE
Mshift [5]	0.7958	1.9725	0.1888	14.41
Ncuts [17]	0.7242	2.9061	0.2232	17.15
Jseg [8]	0.7756	2.3217	0.1989	14.40
SpectClust [6]	0.7357	2.6336	0.2469	15.40
GraphBased [10]	0.7139	3.3949	0.1746	16.67
Mscuts [25]	0.7559	2.4701	0.1925	15.10
ROI-Seg [9]	0.7599	2.0072	0.1846	22.45
NormTree [22]	0.7521	2.4954	0.2373	16.30
Our method	0.6902	1.9028	0.1580	19.7984

Global Consistency Error (GCE) [14]: GCE measures the extent to which one segmentation can be viewed as a refinement of the other.

Boundary Displacement Error (BDE) [11]: BDE measures the average displacement error of the boundary pixels between two segmented images.

We notice that, more is greater the PRI measure, better is the segmentation. In the other hand, more are smaller the VoI, GCE and BDE measures, better is the segmentation. The choice of these four measures where motivated by the fact that those measures where used to evaluate some image segmentation algorithms. Thus, we can compare objectively our segmentation method with those given in [12].

We report in table 2, the average scores over the Berkeley database for eight other segmentation methods: Mean Shift(MShift) [5], "Normalized cuts" (NCut) [17], Graph-based Segmentation [10], Spectral clustering((Spect-Clust) [6], multi-scale normalized cut(Mscuts)[25], MSER-based segmentation (ROI-Seg) [9], normalized partitioning tree (NormTree) [22] and the JSEG algorithm (Jseg)[8]. We set the number of regions in our method to 6 for all the database images. We remark that our method achieves the best performance for VoI and GCE measures. This comparison proves that our segmentation method have less errors in terms of variation of information with respect to the ground truth.

5 Conclusion

In this work, we proposed an hybrid segmentation based on efficient graph based segmentation and normalized cuts partitioning algorithm. The proposed method segment an image with the efficient graph based method, computes the region adjacency graph from the segmented image and finally partition the graph to have the final segmentation. This approach has shown good result with a low computational time.

References

1. Graph based image segmentation tutorial (November 21, 2007), http://www.cis.upenn.edu/~jshi/graphtutorial/
2. Adamek, T., Connor, E., Murphy, N.: Region-based segmentation of images using syntactic visual features. In: 6th International Workshop on Image Analysis for Multimedia Interactive Services (April 2005)
3. Berkeley: Berkeley segmentation and boundary detection benchmark and dataset (2003), http://www.cs.berkeley.edu/projects/vision/grouping/segbench
4. Chen, T.W., Chen, Y.L., Chien, S.Y.: Fast image segmentation based on k-means clustering with histograms in hsv color space. In: Multimedia Signal Processing, pp. 322–325 (2008)
5. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 603–619 (2002)
6. Cour, T., Benezit, F., Shi, J.: Spectral segmentation with multiscale graph decomposition. In: Computer Vision and Pattern Recognition, CVPR (2005)
7. Cour, T., Yu, S., Shi, J.: Normalized cuts matlab code, http://www.cis.upenn.edu/~jshi/software
8. Deng, Y., Manjunath, B.S.: Unsupervised segmentation of color-texture regions in images and video. Pattern Analysis and Machine Intelligence 35, 800–810 (2001)

9. Donoser, M., Bischof, H.: Roi-seg: Unsupervised color segmentation by combining differently focused sub results. In: Computer Vision and Pattern Recognition (CVPR) (2007)
10. Felzenszwalb, P., Huttenlocher, P.: Efficient graph-based image segmentation. International Journal of Computer Vision 59, 167–181 (2004)
11. Freixenet, J., Muñoz, X., Raba, D., Martí, J., Cufí, X.: Yet another survey on image segmentation: Region and boundary information integration. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2352, pp. 408–422. Springer, Heidelberg (2002)
12. Hoon, T., Mu, K., Uk Lee, S.: Learning full pairwise affinities for spectral segmentation. In: Computer Vision and Pattern Recognition (CVPR), pp. 2101–2108 (2010)
13. Makrogiannis, S., Economou, G., Fotopoulos, S.: A region dissimilarity relation that combines feature-space and spatial information for color image segmentation. IEEE Transactions Systems, Man, Cybernetics Part B 35, 44–53 (2005)
14. Martin, D., Fowlkes, C., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. Technical report, EECS Department. University of California, Berkeley, Janvier (2001)
15. Maxwell, A., Shafer, A.: Physics-based segmentation of complex objects using multiple hypotheses of image formation. Computer Vision And Image Understanding 65, 269–295 (1997)
16. Meila, M.: Comparing clusterings by the variation of information. Journal of Multivariate Analysis, 173–187 (2003)
17. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transaction on Pattern Analysis and Maching Intelligence 22, 888–905 (2000)
18. Tao, W., Jin, H., Zhang, Y.: Color image segmentation based on mean shift and normalized cuts. IEEE Transactions on, Systems and Cybernetics-Part B 37, 1382–1388 (2007)
19. Tatiraju, S., Mehta, A.: Image segmentation using k-means clustering, em and normalized cuts. Technical report
20. Thiran, J., Warscotte, V., Macq, B.: A queue-based region growing algorithm for accurate segmentation of multi-dimensional digital images. Signal Processing 60, 1–10 (1997)
21. Unnikrishnan, R., Pantofaru, C., Hebert, M.: Toward objective evaluation of image segmentation algorithms. Pattern Analysis and Machine Intelligence 29, 929–944 (2007)
22. Wang, J., Jia, Y.: Normalized tree partitioning for image segmentation. In: Computer Vision and Pattern Recognition, CVPR (2008)
23. Weiss, Y.: Segmentation using eigenvectors: A unifying view. In: Seventh International Conference on Computer Vision (ICCV 1999), vol. 2, p. 975 (1999)
24. Wu, Z., Leahy, R.: An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 15, 1101–1113 (1993)
25. Yu, S., Shi, J.: Multiclass spectral clustering. In: International Conference on Computer Vision (ICCV), pp. 313–319 (2003)

Stability Analysis of Static Signatures for Automatic Signature Verification

Donato Impedovo and Giuseppe Pirlo

Dipartimento di Informatica, Università degli Studi di Bari "A. Moro", via Orabona 4,
70125 - Bari, Italy
Centro Interfacoltà “Rete Puglia“, Università degli Studi di Bari "A. Moro",
via Giulio Petroni 15F/1, 70100 - Bari, Italy
impedovo@deemail.poliba.it

Abstract. This paper presents a static signature verification system based on the concept of local stability. Stable regions are detected in the signatures, during the enrolling phase, and are considered to be those regions affected by low variations of features among the training set. The stability evaluation is based on the Hamming distance. Stable regions are successively used for verification, in the running phase. A region-oriented verification strategy is considered, based on a well-defined similarity measure which takes into account the variability in signing of the writer. The experimental results, carried out on signatures from the GPDS database, demonstrate the viability of proposed approach.

Keywords: Biometry, Static Signature Verification, Stability Analysis.

1 Introduction

Handwritten signature is a biometric trait with peculiar and interesting characteristics for verification aims. In fact, even if handwritten signature is considered to be a behavioral biometric trait, it depends from physical and psychological conditions of the writer, as well as on the writing device. It follows that automatic signature verification involves aspects from disciplines ranging from human anatomy to engineering, from neuroscience to computer science. Notwithstanding, handwritten signature is well-accepted by users and well-recognized by public and private institutions. Thus, the field of automatic signature verification is actually attracting more and more researchers, interested to both scientific and commercial aspects [1, 2, 3].

Two categories of signature verification systems can be considered, depending on the data acquisition method [1]: static (off-line) systems and dynamic (on-line) systems. Static systems perform data acquisition after the writing process has been completed. In this case, the signature is represented by a grey level image $\{I(u,v)\}_{0\le u\le U,\ 0\le v\le V}$, where I(u,v) denotes the grey level at the position (u,v) of the image. Static systems involve the treatment of the spatio-luminance representation of a signature image. Dynamic systems perform data acquisition during the writing process. In this case on-line acquisition devices are adopted and a signature is

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 241–247, 2011.

represented as a sequence $\{S(n)\}_{n=0,1...N}$, where S(n) is the signal value sampled at time n·Δt of the signing process (0 ≤ n ≤ N), Δt being the sampling period. Thus, dynamic systems concern the treatment of a spatio-temporal representation of the signature [3]. Therefore, no direct dynamic information is available on the signing process when static signatures are considered [4, 5]. Notwithstanding, static signature verification is very important for many application fields, like automatic bank-check processing, insurance form processing, document validation and so on [6, 7, 8].

In order to improve signature verification performance, this paper presents an attempt to extract useful information derived from the signing process, that can be used for signature verification purposes. The main consideration is that diverse regions of a signature convey different amounts of distinctive information. In particular, stable regions can be considered to better represent the user and at the same time, moreover since their discovering is hidden in the training phase of the system which performs the evaluation among a set of training genuine samples, these regions could be more difficult to imitate than others and imperfections of the forgers could be much more easily detectable. Thus signature verification can be performed more effectively if stable regions are considered [9, 10, 11, 12].

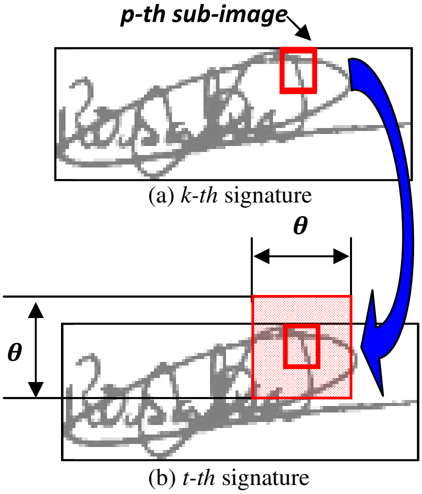
Starting from this consideration, in this paper a technique for the analysis of local stability is used to detect stable regions in a static signature. A signature verification system is then presented which performs signature verification using only stable regions for verification. The organization of the paper is the following. Section 2 presents the stability analysis technique. Section 3 presents the static signature verification system. Section 4 reports the experimental results, obtained on the GPDS database. The conclusion is reported in Section 5.

2 Analysis of Stability in Static Signatures

It is common experience that not all parts of a signature are equally distinctive for signature verification, as widely discussed in recent literature [13, 14, 15]. In order to select distinctive parts of a static signature, in this paper regions on the upper and lower contours of signature are considered (at a first stage), since it is well known that upper and lower contours of a signature can convey relevant information for verification aims. Therefore, after the preprocessing phase, in which each signature is binarized and normalized to a fixed rectangular area, stable regions are detected. More precisely, let $I=I(u, v)$ be a signature image, $0 \leq u \leq U^{max}$, $0 \leq v \leq V^{max}$; $I_{x,y} = I(x+r,y+s)$ a sub-image of size (2R +1, 2S+1) (R,S integers) with $-R \leq r \leq R$, $-S \leq s \leq S$. Furthermore, let $I^1, I^2, \ldots, I^k, \ldots I^K$ be a set of K genuine signature images and $I^k_{x^k_p,y^k_p}$ be the p-th region extracted from I^k. The stability $S\left(I^k_{x^k_p,y^k_p}\right)$ of the region $I^k_{x^k_p,y^k_p}$ is defined as the mean value of the distance between $I^k_{x^k_p,y^k_p}$ and the corresponding regions on the other genuine signatures [16]:

$$S\left(I^k_{x^k_p,y^k_p}\right) = mean\left\{D^*\left(I^k_{x^k_p,y^k_p}, I^t_{x^t_p,y^t_p}\right) | t = 1,2,\ldots,K, k \neq t\right\} \quad (1)$$

where $D^*(I^q_{x^q_p,y^q_p}, I^t_{x^t_p,y^t_p}) = min\{D(I^q_{x^q_p,y^q_p}, I^t_{x^t_p,y^t_p}) \mid |x^q_p - x^t_p| < \theta, |y^q_p - y^t_p| < \theta\}$, being θ a threshold and D(·) the Hamming Distance among the sub-images. Figure 1 shows the matching procedure, in which the *p-th* sub-image of the *k-th* signature is matched against the corresponding sub-image of the *t-th* signature, determined within a wide region. Of course, this procedure allows to assign to each sub-region a degree of stability. Regions that are perfectly replicated by the signer on different genuine signatures will have a stability value (distance value) equal to zero. Conversely, as the sub-region matches worse with the corresponding sub-regions, the distance augments. This means that the signer is not stable in affixing that part of the signature. The stability parameter $S(\cdot)$ ranges in the interval [0,1] and the boundary 0 and 1 represents, respectively, the highest and the lowest degree of similarity.



(a) *k-th* signature

(b) *t-th* signature

Fig. 1. Example of Matching Procedure

3 Stability-Based Signature Verification

The signature verification system consists of four main modules. The Data Acquisition Module (DAM) allows data acquisition of static signatures. The Preprocessing Module (PM) performs signature binarization, signature normalization to a fixed rectangular area and noise reduction. The Feature Extraction Module (FEM) performs the extraction of the discriminative features. More precisely, in the enrollment phase the stable regions of the signature are extracted, according to the approach described in Section 2. The Classification Module (CM) performs signature verification according to a two-level strategy [15]: first it applies a matching rule on stable regions, based on a simple similarity measure, to derive verification responses at the region-level; successively, local decisions are combined to obtain the final decision at the signature level. In the first stage the stable regions on the unknown signature are individually verified. For this purpose, each region is matched against the corresponding regions on the genuine samples. In the regional matching stage a set of N sub-images are extracted from the test signature image and compared against N corresponding stable sub-images on the reference signature images. Let I^t be the image of a test signature and I^1, I^2,… ,I^k,… I^K be a set of K reference signature images. Furthermore, let $I^t_{x^t_1,y^t_1}, I^t_{x^t_2,y^t_2}, \dots, I^t_{x^t_N,y^t_N}$ be N sub-images of I^t selected for signature matching and corresponding to stable regions detected in the training phase. A sub-image $I^t_{x^t_n,y^t_n}$ of the test signature is considered as belonging to a genuine signature if and only if

$$\left|\Delta(I^t_{x^t_n,y^t_n}) - \mu_n\right| \le \sigma_n \qquad (2)$$

where:

$$\Delta(I^t_{x^t_n,y^t_n}) = \min_{k} \min_{(x^k_n,\, y^k_n)} D(I^t_{x^t_n,y^t_n}, I^k_{x^k_n,y^k_n}) \qquad (3)$$

with $|x^k_p - x^v_p| < \theta$ and $|y^k_p - y^v_p| < \theta$, μ_n and σ_n are the mean and the standard deviation computed for the set of stable sub-images $I^1_{x^1_1,y^1_1}$, $I^2_{x^2_2,y^2_2}$, …., $I^k_{x^k_n,y^k_n}$, …., $I^K_{x^K_N,y^K_N}$ of the genuine signatures.

In the second stage a majority voting strategy is used to combine local decisions and produce the final verification decision, the signature is considered to be genuine if and only if $R_g \ge \lambda_s \cdot R_{tot}$, where R_g is the number of genuine sub-regions, R_{tot} is the total number of sub-regions considered for signature verification and λ_s is a personal threshold value of the signer which estimates his/her personal variability in signing.

4 Experimental Results

The GPDS database has been considered: 16200 signatures from 300 individuals have been used. In particular, for each individual, there are 24 genuine signatures and 30 forgeries [17]. Each signature was normalized to a box of 113x43 pixels, the sub-regions inspected in order to evaluate their stability have a size of 11x11 pixels. For each signer the most stable regions are identified according to the procedure described in Section 2, and at a first stage, the number of stable regions selected by the described approach range from 3 to 6 for each signature.

Results are reported in table 1 in terms of Type I error rate (i.e. False Rejection Rate - FRR) and of Type II error rate (i.e. False Acceptance Rate - FAR). Moreover FAR is evaluated both considering skilled and random forgeries. The first, second and third rows refer, respectively, to the use of regions selected on the lower, upper and *lower.and.upper* contour of the signature.

Table 1. Error rate

	FRR	FAR-skilled	FAR-random
Lower Contour	40%	31%	17%
Upper Contour	22%	25%	13%
Lower+Upper Contour	25%	26%	22%

One of the biggest problem, due to the use of the upper and lower contours in order to investigate stable regions, is related to the simplicity and linearity of the contour. In fact if, for instance, the lower contour is an underlining trait, each region on that contour will result in a very high stability (low distance), but at the same time it will be not discriminative. To partially overcame this problem, the stability of each portion of the signature could be evaluated. Figure 2 shows the stability of regions of two static signatures: high stability refers to regions showing little distance values. It can be observed that circular and straight traits have an high degree of stability.

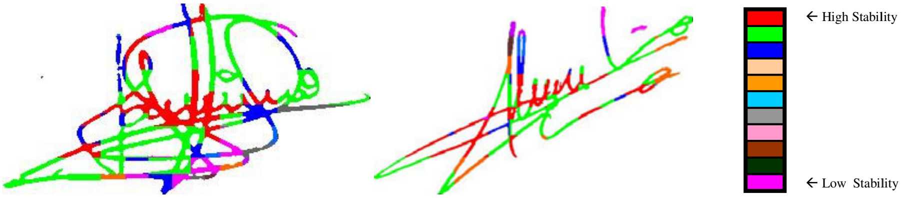


Fig. 2. The stability is evaluated by taking into account each portion of the signature

However, since it is well known that, in general, the upper and the lower contour can provide useful features for signature matching and in order to reduce the computational load due to the inspection of each portion of the signature, the inspection of regions located in middle part of the signature is performed when the upper (lower) part of the profile containing the stable regions is evaluated to be a linear or a circular one. To this aim, the trajectory of the upper (lower) profile has been evaluated by considering the derivatives of traits on consecutive windows. According to this procedure, if the region is considered to convey not discriminative features, the middle part of the signature is inspected (figure 3).

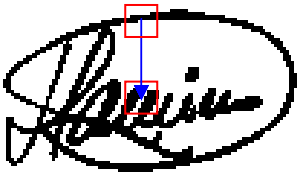


Fig. 3. Skipping not discriminative traits

Under these assumptions the Type I error rate is 20%, whereas the type II error rates for random and skilled forgeries is respectively 19% and 22%. These results are comparable with other results on the same database carried out in the recent literature [18].

It must be underlined that a user-dependent threshold can be adopted to select specific stable regions starting from $S(\cdot)$ *(1)*. The evaluation of $S(\cdot)$ could be improved by comparing the ability of forgery in replicating the specific region under analysis. This should be taken into account in future works.

5 Conclusion

This paper presents a new technique for the analysis of stability in static signatures, based on Hamming Distance. In particular stable regions are detected by comparing multiple genuine signatures. These regions are then considered for automatic signature verification and a local verification decision is derived for each region. A Majority Vote schema is finally used to combine decisions achieved at regional level. The experimental results, carried out on signatures from the GPDS database, demonstrate that the proposed approach can be considered as a first implementation of a local stability technique. Of course the approach needs more research, in fact the definition of stability must take into account also the discrimination capability of the specific region, to this aim forgeries samples could be considered in the training phase.

References

1. Plamondon, R., Lorette, G.: Automatic Signature Verification and Writer Identification – The State of the Art. Pattern Recognition 22(2), 107–131 (1989)
2. Leclerc, F., Plamondon, R.: Automatic Signature Verification: The State of the Art – 1989 1993. In: Plamondon, R. (ed.) IJPRAI, vol. 8(3), pp. 643–660. World Scientific, Singapore (1994)
3. Impedovo, D., Pirlo, G.: Automatic Signature Verification – The State of the Art. IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Review 38(5), 609–635 (2008)
4. Plamondon, R., Srihari, S.N.: On line and Off line Handwriting Recognition: A Comprehensive Survey. IEEE T-PAMI 22(1), 63–84 (2000)
5. Dimauro, G., Impedovo, S., Lucchese, M.G., Modugno, R., Pirlo, G.: Recent Advancements in Automatic Signature Verification. In: 9th International Workshop on Frontiers in Handwriting Recognition (IWFHR-9), Kichijoji, October 25-29, pp. 179–184 (2004)
6. Dimauro, G., Impedovo, S., Pirlo, G., Salzo, A.: A multi-expert signature verification system for bankcheck processing. International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI) 11(5), 827–844 (1997)
7. Yoshimura, M., Yoshimura, I.: Investigation of a verification system for Japanese countersignatures on traveler's checks. Trans. IEICE J80-D-II(7), 1764–1773 (1997)
8. Lee, L.L., Lizarraga, M.G., Gomes, N.R., Koerich, A.L.: A prototype for Brazilian bankcheck recognition. IJPRAI 11(4), 549–569 (1997)
9. Ferrer, M.A., Alonso, J.B., Travieso, C.M.: Offline Geometric Parameters for Automatic Signature Verification Using Fixed-Point Arithmetic. IEEE T-PAMI 27(6), 993–997 (2005)
10. Nouboud, F.: Handwritten signature verification: a global approach. In: Impedovo, S. (ed.) Fundamentals in Handwriting Recognition, pp. 455–459. Springer, Heidelberg (1994)
11. Ramesh, V.E., Narasimha Murty, M.: Off-line signature verification using genetically optimized weighted features. Pattern Recognition 32(2), 217–233 (1999)
12. Bajaj, R., Chaudhury, S.: Signature Verification Using Multiple Neural Classifiers. Pattern Recognition 30(1), 1–7 (1997)

13. Congedo, G., Dimauro, G., Forte, A.M., Impedovo, S., Pirlo, G.: Selecting Reference Signatures for On Line Signature Verification. In: Braccini, C., Vernazza, G., DeFloriani, L. (eds.) ICIAP 1995. LNCS, vol. 974, pp. 521–526. Springer, Heidelberg (1995)
14. Impedovo, D., Modugno, R., Pirlo, G., Stasolla, E.: Handwritten Signature Verification by Multiple Reference Sets. In: Proc. of the 11th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 19–21 (August 2008)
15. Impedovo, D., Pirlo, G.: On the Measurement of Local Stability of Handwriting - An application to Static Signature Verification. In: Proc. of Biometric Measurements and Systems for Security and Medical Applications (BIOMS 2010), Taranto, Italy, pp. 41–44. IEEE Computer Society Press, Los Alamitos (2010)
16. Pirlo, G., Impedovo, D., Stasolla, E., Trullo, C.A.: Learning local correspondences for static signature verification. In: Serra, R., Cucchiara, R. (eds.) AI*IA 2009. LNCS, vol. 5883, pp. 385–394. Springer, Heidelberg (2009)
17. Vargas, J.F., Ferrer, M.A., Travieso, C.M., Alonso, J.B.: Off-line Handwritten Signature GPDS-960 Corpus. In: Proc. 9th ICDAR, September 23-26, vol. 2, pp. 764–768 (2007)
18. Jayadevan, R., Kolhe, S.R., Patil, P.M.: Dynamic Time Warping Based Static Hnad Printed Signature Verification. Journal of Pattern Recognition Research 4(1), 52–65 (2009)

Segmentation Strategy of Handwritten Connected Digits (SSHCD)

Abdeldjalil Gattal[1,2] and Youcef Chibani[3]

[1] Université de Tébessa, Algeria
[2] Ecole Nationale Supérieure d'Informatique (ESI), Oued Smar, Algeria
ab.gattal@gmail.com
[3] Laboratoire de Communication Parlée et Traitement des Signal,
Faculté d' Electronique et d'Informatique, University of Sciences and
Technology Houari Boumedienne, Bab-Ezzouar, Algiers, Algeria
ychibani@usthb.dz

Abstract. The handwritten digit segmentation is the most important module for handwritten digit recognition, which constitutes a difficult task because of overlapping and / or connected of adjacent digits. To resolve this problem, several segmentation methods have been developed each one having its advantage and disadvantage. In this work, we propose a segmentation approach depending of the configuration link between digits. With the help of a few rules, multiple hypotheses are defined for finding the best segmentation path in order to separate two connected digits. Hence, a verification strategy is proposed in order to generate all possible segmentation-recognition hypotheses. The performance of our strategy is evaluated in terms of correct recognition rates using the confusion matrix.

Keywords: recognition, segmentation, segmentation-recognition, handwritten digits, verification strategy, Support Vector Machines.

1 Introduction

The handwriting recognition is interested in many applications such as automatic sorting postal mail, the automatic processing of the administrative files, or the recording the courtesy amount of the bank checks [1]. In this work, we are interested to develop a segmentation method for finding the best way of separation when the digits are connected.

Usually, the handwritten digits recognition can be done mainly in three steps. The first step facilitates further processing (smoothing or reorganization of characters, homogenization of the line thickness, etc...) in order to locate and isolate the characters more easily from each other. This step, called segmentation, [2],[3],[4],[5] consists in separating the text into elementary characters for which the limited number of characters defines the possible distinct classes. Then, a feature generation is performed on the character image for reducing the dimension of the representation and thus makes the design of the classification system. Finally, a decision function allows assigning a character image to predefined class.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 248–254, 2011.

Such recognition system, the segmentation constitutes the most difficult step, which is related to several factors such as the slope of the figures, overlapping of two digits, and the joining of two consecutive digits or the default inking.

Usually, the segmentation can be conducted in considering three following situations: distinct digits and overlapped digits or connected digits. In most cases, the overlapped and connected digits are the frequent situations observed. Hence, many algorithms were proposed to separate the couples of contiguous digits. Some ones are based on contours and others on the skeleton or on the size, the number and the position of the water reservoir to deduce the potential points of cutting [6], [1], [3].

In this paper, we only are interested in the segmentation where we try to improve the performances by offer the best cutting when the digits are naturally connected. Hence, we propose a strategy based on an isolated verifier, which was responsible for detecting both over-segmentation and under-segmentation for solving the majority of the segmentation problems [8].

The paper is organized as follows. In section 2, we review the segmentation method and its strategy. The section 3 is devoted to present the experimental results. Finally, the conclusion and future work are presented in Section 4.

2 Segmentation Method

The segmentation consists to find the best way of cutting to isolate the digits [6],[7],[9]. However, finding the best way is not straightforward because two contiguous digits can be insulated, deformed, related by one or more contact points, partially superimposed, or on the contrary written in several pieces. Moreover, several factors such as the variability of the style and the tool of writing increase the complexity of segmentation. Fig.1 illustrates some difficult examples.

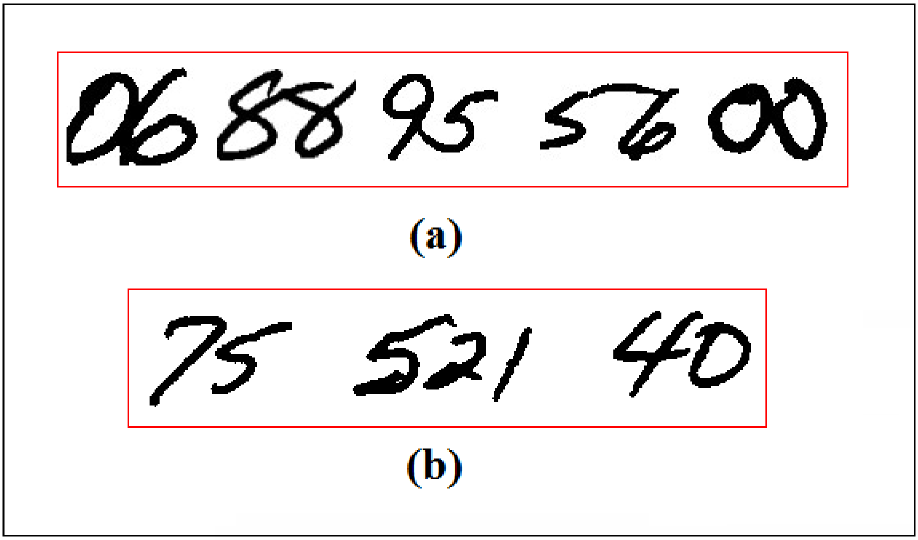


Fig. 1. Some difficult examples

(a) Connected digits (b) Overlapped digits

In this paper, we propose an approach based on the verifications strategy in order to generate all possible segmentation-recognition hypotheses. In the following, we review how to find all possible segmentations paths, by analyzing the interconnection points [2].

2.1 Segmentation Based on the Interconnection Points

This method has been proposed more recently which involves analyzing the number and nature of interconnection points between two adjacent digits in order to define the optimal position for cutting a digit image couple [2]. The first step is to define the Interconnection Points (IPs) and the Bases Points (BPs) from which to start the segmentation. BPs are obtained from the extrema (minima and maxima) detected on the local contour connected components while IPs are calculated using the Freeman code according to the 8 directions in the clock-wise. Often, the connection points contain IPs, BPs or both at the same time. Hence, three hypotheses can be considered for the optimal segmentation:

- Hypothesis 1: If the Euclidean distance between the projection of the BP and the IP is lower than a threshold, the cut is made in the vicinity of the IP (Fig.2.a).
- Hypothesis 2: if the lower segment of IP is related to a upper segment of IP (or vice versa) and both IPs are near a BP, the skeleton path linking both IPs (Skeleton path) is used as part of the segmentation cut with complementary paths between BPs and IPs (Composed path) (Fig.2.b)
- Hypothesis 3: in some cases, even if there is a connection between two digits, the skeleton path does not have an IP. Thus, to avoid the under-segmentation (lack of segmentation point), the algorithm builds a path of segmentation based minimal Euclidean distance between bases points upper and bases points lower (closest points) in the middle (Fig.2.c).

2.2 Segmentation Strategy of Connected Digits

The main problem for separating two connected digits is the detection of the interconnection points (IP). Hence, we propose the following steps:

- For detecting the IP, we set up a window where the height has the same height as the original image and its width is constant. The IP is located in the middle of this width.
- For each window, if a single IP is detected, the cut is conducted in the vicinity of IP when the Euclidian distance between IP and BP (BP upper or BP lower) is lower than a threshold T1. If there are two IPs (IP upper and IP lower), then we apply the hypothesis 2, in order to separate the two digits. To avoid under-segmentation, we apply the Hypothesis 3 in all cases.

The threshold T1 is determined through experimentation, seeking the maximum Euclidean distance between IP and BP in the case of connected digits.

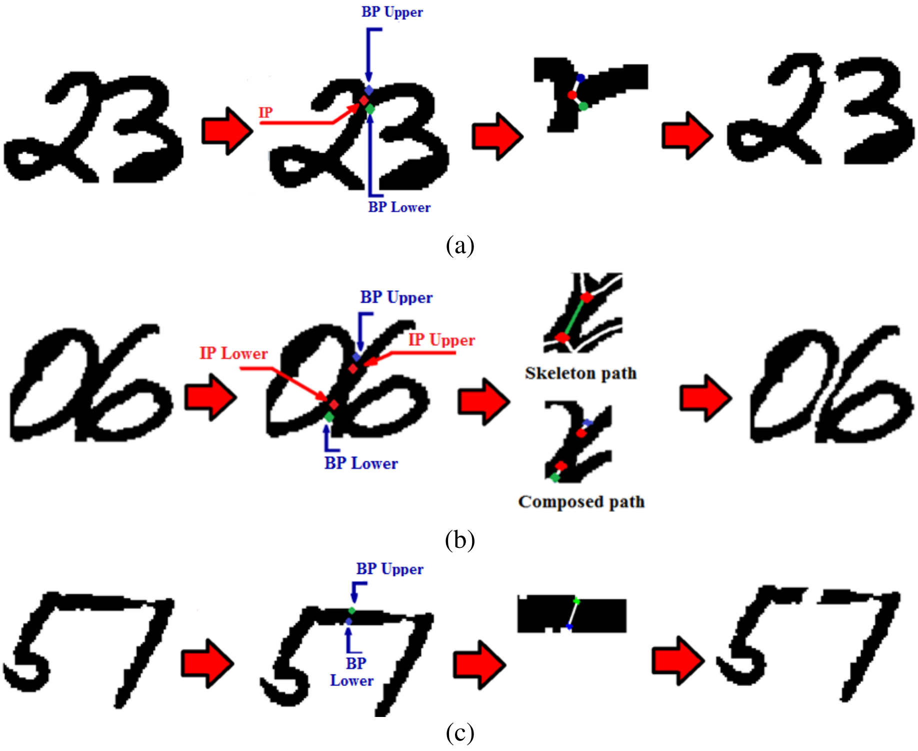


Fig. 2. Segmentation paths according to IP and BP positions collected

(a) Hypothesis 1 (b) Hypothesis 2 (3) Hypothesis 3

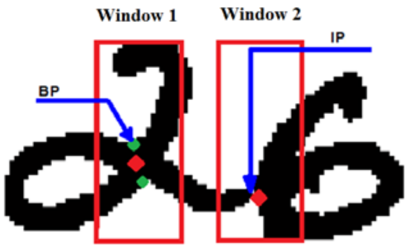


Fig. 3. Set up windows

2.3 Recognition and Verification

The final decision for a segmentation-recognition hypothesis is provided by the average of the decision functions made by its sub-components. The decision function of a sub-component is provided by the SVM classifier.

In Figure 4, we present an example of segmentation-recognition. To better illustrate this, we took a better result than the correct one. This result takes the maximum value of the averages decision functions between the results of the segmentation hypothesis. This verification allows reducing the confusion between isolated and under-segmented characters.The diagram in Figure 4 is designed completely segmentation, recognition and verification. In the column "outputs", we can see the results of the decision function produced, for each SVM classifier [11] according to the rules (hypotheses) presented in Figure 3.

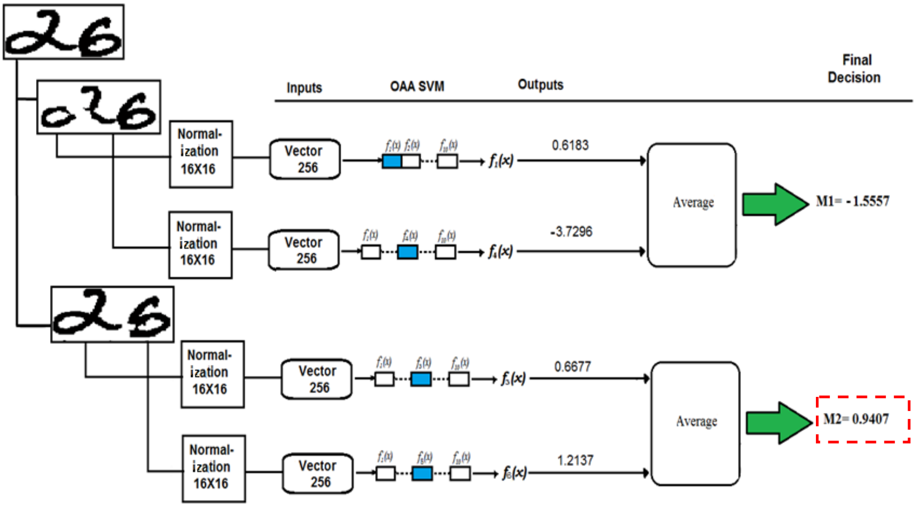


Fig. 4. Final decision for the segmentation-recognition

3 Experimental Results

The evaluation of a segmentation method is very subjective. It can be done in two steps. The first is to evaluate the segmentation based on a priori knowledge about the effects of segmentation. The second step is to evaluate the system for recognizing handwritten digits by integrating segmentation and classification. In our case, we use the system evaluation method for appreciate the quality of the segmentation on a database of NIST SD19 [10].

The NIST SD19 database is divided into two parts: the first consisting of 5000 digits is used for learning and the second 500 digits is used for testing. The test database contains 250 digits linked. Each digit is normalized to 16x16 pixels after segmentation.

The recognition module is based on a SVM classifier with RBF kernel. The multi-class SVM implementation is based on the approach "one against all". It involves using a binary classifier by category. The SVM and RBF parameters are fixed to C=10 and σ =10, respectively. Furthermore, the threshold T1 is fixed experimentally to 7.

The evaluation is conducted without feature extraction in order to evaluate the quality of segmentation. This evaluation is illustrated in the confusion matrix reported in Table 1.

Table 1. Recognition rates obtained for each class

		Reference Classes										
		0	1	2	3	4	5	6	7	8	9	
Computed Classes	0	81.08	2.70	2.70	2.70	2.70		2.70		5.41		
	1		100.00									
	2			77.27					9.09	4.55		
	3				66.67					4.76		
	4					81.25	6.25					
	5						91.18	8.82				
	6			6.25				81.25				
	7								100			
	8									88.89		
	9					5.00					95.00	
	Overall Rate of SSCHD method.											86.26

We can note that some digits are not recognized, which are considered as rejected by the SVM classifier. Furthermore, we can note that some digits are difficult to recognize specifically the couples ('0','8'), ('2','7') and ('3','8'). These digits are more problematic for recognition since they are sometimes visually very similar. The decision is difficult even for a human operator.

4 Conclusion and Future Work

The objective of this paper is the presentation of a strategy for segmenting connected handwritten digits, in order to find the best cut to isolate two adjacent digit images.

The method allows improving the performances and resolving many problems of connected digits. It uses conjointly segmentation-verification for finding the way of cutting of the connected digits.

The first results obtained are encouraging since we can manage to follow, with a weak error of detection, all the possible ways of segmentation. This combination uses few rules and has the advantage of providing a correct segmentation in the most cases.

For future work, we try to complete the system by adding an analysis module for generating the features. For a complete validation of the system, we also try to evaluate the system on a larger database.

References

1. Dimauro, G., Impedovo, S., Pirlo, G., Salzo, A.: Automatic Bankcheck processing: A New Engineered System. International Journal of Pattern Recognition and Artificial Intelligence 11(4), 467–504 (1997)
2. Vellasques, E., Oliveira, L.S., Britto Jr., A.S., Koerich, A.L., Sabourin, R.: Filtering segmentation cuts for digit string recognition. Pattern Recognition 41(10), 3044–3053 (2008)

3. Congedo, G., Dimauro, G., Impedovo, S., Pirlo, G.: Segmentation of Numeric Strings. In: Proc. of Third Int. Conf. on Document Analysis and Recognition, Canada, pp. 1038–1041. IEEE Computer Society, Montreal (1995)
4. Jang, B.K., Chin, R.T.: One-pass parallel thinning: Analysis, properties, and quantitative evaluation. IEEE Transactions on Pattern Analysis and Machine Intelligence 14(11), 1129–1140 (1992)
5. Shridhar, M., Badreldin, A.: Recognition of Isolated and Simply Connected Handwritten Numerals. Journal of Pattern Recognition 19(1), 1–12 (1986)
6. Ayat, N.E., Cheriet, M., Suen, C.Y.: Un système neuro-flou pour la reconnaissance de montants numériques de chèques arabes. Colloque international francophone sur l'écrit et le document, Montréal, Québec, Canada, pp. 03–07 (2000)
7. Hussein, K.M., Agarwal, A., Gupta, A., Wang, P.S.P.: A knowledge-based algorithm for enhanced recognition of handwritten courtesy amounts. Pattern Recognition 32, 305–316 (1999)
8. Oliveira, L.S., Sabourin, R., Bortolozzi, F., Suen, C.Y.: A modular system to recognize numerical amounts on Brazilian bank cheques. In: Proc. of 6th International Conference on Document Analysis and Recognition (ICDAR), Seattle, USA, pp. 389–394. IEEE Computer Society, Los Alamitos (2001)
9. Fujisawa, H., Nakano, Y., Kurino, K.: Segmentation Methods for Character Recognition: From Segmentation to Document Structure Analysis. Proceedings of the IEEE 80(7), 21–28, 1079–1091 (1996)
10. Grother. P.J.: NIST Special Database 19; Handprinted Forms and Characters Database. National Institute of Standards and Technology, NIST (1995)
11. Schölkopf, B., Burges, C., Vapnik, V.: Extracting support data for a given task. In: KDD 1995, pp. 252–257 (1995)

An Experimental Comparison of Different Methods for Combining Biometric Identification Systems

Emanuela Marasco and Carlo Sansone

Dipartimento di Informatica e Sistemistica,
Università degli Studi di Napoli Federico II
Via Claudio, 21 I-80125 Napoli, Italy
{emanuela.marasco,carlosan}@unina.it

Abstract. Several works in the recent literature on biometrics demonstrate the efficiency of the multimodal fusion to enhance performance and reliability of the automatic recognition. In this paper, we experimentally compare the behavior of different rules for integrating different biometric identification systems. We investigated how the benefits of the fusion change by varying the set of the fused modalities, the adopted fusion scheme and the performance of the individual matchers. The experiments were carried out on two multimodal databases, using face and fingerprint. We considered trained and fixed fusion methods at score, rank and decision level.

1 Introduction

In the recent literature on biometrics, several researches demonstrate the efficiency of the multimodal fusion to enhance performance and reliability of the automatic recognition [1]. Integrating biometric information from multiple sources, multimodal biometric systems are able to improve the authentication performance, increase the population coverage, offer user choice, make biometric authentication systems more reliable and robust to spoofing.

However, the benefits of multibiometrics depend on the accuracy, complementarity, reliability and quality measurement of their component biometric experts. Moreover, when designing a multibiometric system, several factors should be considered. These concern the choice and the number of biometric traits, the level of integration and the mechanism adopted to consolidate the information provided by multiple traits. Fusion at match score level is usually preferred due to the easy to access and combine the scores presented by different modalities. The parallel fusion strategy has been extensively explored, however serial and hybrid architectures present important advantages. In particular, the serial fusion considers the biometric matchers one at a time, and makes a reliable decision by employing few experts and activating the remaining experts only for difficult cases. In general, it is desirable that a fusion scheme involves statistically independent modality matchers. In a multimodal fusion, the set of expert outputs

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 255–264, 2011.

is expected to be statistically independent, while in intramodal fusion, where the component matchers rely on the same biometric trait, a high dependency is expected among the expert outputs [2]. The merit of both multimodal and intramodal fusion has been demonstrated in [3].

Moreover, although individual modalities have proven to be reliable in ideal environments, they can be very sensitive to real environmental conditions. In real scenarios, it is difficult to acquire high quality samples, then biometric authentication errors are inevitable. The impact of adverse environmental conditions on the characteristics of the collected biometric data can be quantified by *quality measures*. It is evident that, a degradation in the quality level of the biometric signal input may affect the reliability of the matching process. The performance of the single modality matcher may change as the data quality changes and different modality matchers are sensitive to different aspects of the signal quality. Then, the opinion of a matcher in the decision of the ensemble have to be appropriately weighted, by assigning a higher weight to the matcher with higher quality data. The same observation has to be considered for the reliability, accuracy and competence of each component matcher [4]. The effectiveness of using quality measures in the fusion has been demonstrated in [5].

The key to create a secure multimodal biometric system is in how the information from different modalities is fused [1]. In identification mode, the consolidation of biometric information can be performed at various levels: sensor level, feature extraction level, match score level, rank level and decision level. Consolidating data at an early stage of the recognition process involves a higher informative contain concerning the biometric input. Thus, it is potentially able to provide better recognition results, but in practice concatenating data at a level before matching may result difficult or not possible. In particular, images captured from sensors with a different resolution are not compatible and feature vectors may be not accessible (they often are proprietary). Combining match scores provided from different matchers is the most effective fusion strategy because they offer the best trade-off between the informative contain and the ease to implement the fusion.

In this paper, we experimentally compare the behavior of different integration rules at different fusion levels for biometric identification systems. Considering the state-of-the-art this kind of comparison has been done only for the verification task and not for the identification [6]. We investigated how the benefits of the fusion change by varying the set of the fused modalities, the fusion scheme exploited and the performance of the individual matchers. The paper is organized as follows. Section 2, presents the analyzed combination rules. Section 3 describes the data, the experimental procedure and it reports our results. Section 4 draws our conclusions.

2 The Considered Fusion Methods

The comparison was carried out by considering the most commonly adopted fusion mechanisms for identification scenario. Among the possible levels, we focus on fusion approaches at score, rank and hybrid rank-score level.

2.1 Fusion Approaches at Score-Level

Fusion at match score level concerns combining the match scores generated by multiple classifiers in order to make a decision about the identity of the subject. In literature, the fusion at score level is performed by employing different approaches [7] based on different models [8]. We considered the *transformation-based schemes* which are described below.

The match scores provided by different matchers are firstly transformed into a common domain (*score normalization*) which refers to changing the location and scale parameters of the match score distributions outputs of the individual matchers [9]. Then, normalized match scores are combined using a simple fusion rule. The operators which are commonly used in the literature are *min*, *max*, *median*,*weighted sum* and *weighted product*, defined by (1), (2), (3), (4) and (5).

$$s_{min} = \min_k s_k \tag{1}$$

$$s_{max} = \max_k s_k \tag{2}$$

$$s_{median} = median_k s_k \tag{3}$$

$$s_{sum} = \sum_{k=1}^{K} w_k s_k \tag{4}$$

$$s_{prod} = \prod_{k=1}^{K} s_k^{wk} \tag{5}$$

where w_k are parameters that need to be estimated. The simple *sum* operator (or *mean*) is a special case of *weighted sum* with $w = \frac{1}{N}$, while the *product* operator is a special case of *weighted product* with $w = 1$. The operators which do not contain parameters to be tuned, are known as *fixed* combiners [10]. Based on experimental results, researchers agree that *fixed* rules usually perform well for ensemble of classifiers having similar performance, while *trained* rules handle better matchers having different accuracy. Thus, when fusing different modalities, individual matchers often exhibit different performance, then for this problem *trained* rules should perform better than *fixed* rules [6]. It has been shown that, the simple sum rule gives very good accuracy in combining multiple biometric systems [6]. Due to the diversity of scenarios encountered in the datasets, training and using a single fusion rule on the entire dataset may not be appropriate. Recently [11], the idea of dynamically selecting biometric fusion algorithms has been adopted.

2.2 Fusion Approaches at Rank-Level

For systems operating in identification mode, rank level fusion is a viable option. It provides a richer information into the decision-making process compared to the decision level, without requiring a normalization phase before combining [12]. Let K be the number of matchers to be fused and N the number of enrolled users. Let r_{ij} be the rank assigned to the j^{th} user enrolled in the database by the i^{th} matcher, $i = 1 \ldots K$, and $j = 1 \ldots N$, then R_{ij}.

Highest rank scheme. For each subject, the combined rank is given by the lowest rank (6). This rank fusion technique presents the advantage of utilizing the strength of each matcher.

$$R_i = \min_{k=1}^{K} r_{ik}, \qquad i = 1, 2, ...N \tag{6}$$

Borda Count scheme. For each subject, the combined rank is given by the sum of the ranks assigned by the individual matchers (7). Such a rule presents the advantage of taking into account the variability of the single matcher outputs. Its drawbacks lie in the assumptions that, the matchers are statistically independent and they perform equally well. This makes the Borda Count method highly vulnerable to the effect of weak classifiers.

$$R_i = \sum_{k=1}^{K} r_{ik}, \qquad i = 1, 2, ...N \tag{7}$$

Logistic regression scheme. The fused rank is a weighted sum of the individual ranks.

$$R_i = \sum_{k=1}^{K} w_k r_{ik}, \qquad i = 1, 2, ...N \tag{8}$$

The weight w_k, $i = 1 \ldots K$, (see equation (8)), is determined through a training phase by logistic regression. This method is useful when the different biometric matchers have significant differences in their accuracies [8].

There is increasing interest in impact of the matcher reliability estimation in the context of fusion in biometrics. However, incorporating reliability information in rank level fusion represents a topic whose the discussion in the literature is at present still limited. The idea is to use reliability in a multibiometric system for reducing the weight of potential incorrect unimodal decisions.

2.3 Fusion Approaches at Hybrid Rank-Score Level

We considered the *predictor-based majority voting*, the *predictor-based sequential* and *predictor-based borda count* proposed in [13]. For each modality, a classifier (predictor) was trained using the hybrid information given by the ratios between scores in terms of ranks with respect to the rank one identity. Such a classifier is used to learn the decision boundary between the correct identification region and the erroneous one.

For a given probe, K unimodal matchers are employed and the winner is the identity to which the majority of matchers have assigned a rank value equal to one. The majority vote will result in an ensemble decision [14]:

$$\arg \max_{i=1 \ldots N} \sum_{k=1}^{K} d_{ik} \cdot v_k \tag{9}$$

where the binary variable d_{ik} is 1 if the k^{th} matcher outputs identity i in rank-1, and the binary variable v_k is 1 if the identification is deemed to be *correct* by the k^{th} predictor. The majority vote scheme assigns an identity to the probe only if the output of at least $\lfloor \Sigma_{k=1}^{K} v_k \rfloor + 1$ unimodal systems correspond to the same identity and are deemed to be correct by v_k.

In the serial scheme, the decisional process is split into two successive stages [15]. The subject to be authenticated submits the first biometric modality to the system which is processed and matched against all the templates present in the gallery. If the resulting identity is labeled to be correct by the predictor module, the input biometric trait is associated to the current identity, otherwise the system suspends the decision and an additional processing stage is performed. In the second stage, *K-1* additional biometric modalities are automatically requested and a voting strategy involving *K-1* unimodal matchers is adopted in the second stage. It can be formulated as follows:

$$Id_m = \begin{cases} Id_u, & \text{if} \quad v_u = 1 \\ \arg\max_{i=1\ldots N} \sum_{k=1}^{K-1} d_{ik} \cdot v_k & \text{if} \quad v_u = 0 \end{cases} \tag{10}$$

where Id_m is the output of the multimodal system and Id_u is the output of the unimodal system at the first stage. In order to maximize the performance of the multimodal system in terms of accuracy and recognition time, on this second stage all the matchers are combined and further stages are avoided.

In the *Borda Count* model, the rank for each identity in the database is calculated as the weighted sum of the individual ranks assigned by the K modality matchers:

$$R_i = \sum_{k=1}^{K} w_k \, r_{ik}, \qquad i = 1, 2, \ldots N \tag{11}$$

In the predictor-based fusion scheme, the unimodal outputs labeled as errors by the predictor have to be excluded from the sum in the equation above which determines the fused rank for each identity. The weight w_k was computed as the ratio between the number of correct identifications detected by the predictor and the total number of test probes.

2.4 Fusion Approaches at Decision-Level

We considered the *pure majority voting* scheme, where the final output corresponds to the most commonly occurring output. For a given probe, the outputs of K modality matchers are examined and the identity to which the majority of matchers have assigned rank one is the *winner*. The majority vote will result in an ensemble decision [14]:

$$\arg \max_{i=1\ldots N} \sum_{k=1}^{K} d_{ik} \tag{12}$$

where the binary variable d_{ik} is 1 if the k^{th} matcher outputs identity i in rank-1.

3 Experimental Results

3.1 Datasets

The performance of the proposed strategy was evaluated on two databases. The first is the West Virginia University (WVU) multimodal biometric database. A subset of this database pertaining to the fingerprint (left thumb [FL1], right thumb [FR1], left index [FL2], right index [FR2]) and face modalities of 240 subjects was used in our experiments. Five samples per subject for each modality were available. Table 1 provides the details of the database. For the *face* modality, frontal images were collected in a controlled scenario. For the *fingerprint* modality, images were collected using an optical biometric scanner, without explicitly controlling the quality [16]. The entire dataset was divided into five sets: the first sample of each identity was used to compose the *gallery* and the remaining four samples of each identity were used as *probes* (P_1, P_2, P_3, P_4). The VeriFinger software was used for generating the fingerprint scores and the VeriLook software was used for generating the face scores.

Table 1. WVU Multimodal Biometric Database

Biometric	Subjects	Samples	Scores
Face	240	5 per subject	Gen 1200 × 4 Imp 240 × 239 × 25
Fingerprint	240	5 per finger	Gen (1200 × 4) × 4 Imp (240 × 239 × 25) × 4

The second database is a subset of the BioSecure multimodal database. This database contains 51 subjects in the Development Set (training) and 156 different subjects in the Evaluation Set (testing). For each subject, four biometric samples are available over two sessions: session 1 and session 2. The first sample of each subject in the first session was used to compose the gallery database while the second sample of the first session and the two samples of the second session were used as probes (P_1, P_2, P_3). For the purpose of this study, we used the face and three fingerprint modalities, denoted as *fnf*, *fo1*, *fo2* and *fo3*, respectively [17]. The details about the number of match scores per person are reported in Tables 2.

3.2 Results

In this paper, we studied the behavior of trained and fixed rules when the combined matchers showed good individual performance and when the matchers achieved a significant individual percentage of errors. Our first experiments focused on comparing the performance of the fusion rules for a set of modality matchers having different a classification capability. Regarding WVU database we used one face matcher and four fingerprint matchers, while regarding Biosecure database we used one face matcher and three fingerprint matchers.

Table 2. The Biosecure database: Development(Dev) and Evaluation(Eva) sets

Dataset	Biometric	Subjects	Samples	Scores
Dev set	Face	51	4 per subject	Gen 204 × 3 Imp 51 × 50 × 16
Dev set	Fingerprint	51	4 per subject	Gen (204 × 3) × 3 Imp (51 × 50 × 16) × 3
Eva set	Face	156	4 per subject	Gen 624 × 3 Imp 156 × 155 × 16
Eva set	Fingerprint	156	4 per subject	Gen (624 × 3) × 3 Imp (156 × 155 × 16) × 3

Tables 3, 4 compare the performance of the existing fusion schemes which are subdivided in fixed and trained rules. WVU data, acquired in optimal environment conditions, offer a scenario where fixed rules are able to achieve good performance. Biosecure data represent a challenging scenario where fixed fusion rules do not perform well, due to the presence of samples with low quality; however, the score sum presents high accuracy, such a scheme represents a good choice to make fusion. The considered trained rules also are able to achieve good multimodal performance.

Table 3. Performance of the analyzed fusion schemes averaged on the four probe sets in the WVU database

Level	Type	Rule	Accuracy
score	fixed	sum	99.58%
score	fixed	min	85.37%
score	fixed	max	99.44%
score	fixed	median	99.26%
score	fixed	product	89.81%
rank	fixed	borda count	95.42%
rank	fixed	highest rank	91.46%
decision	fixed	majority voting	98.75%
score	trained	weighted sum	98.23%
hybrid	trained	predictor-based majority voting	100%
hybrid	trained	predictor-based borda count	96.67%
hybrid	trained	predictor-based sequential	99.59%

Our further experiments aimed to compare the performance of the fusion schemes when increasing the number of the combined modality matchers (see Fig. 1). We proceeded with adding matchers in the fusion scheme based on their performance. On WVU database, the number of combined matchers ranges from

Table 4. Performance of the analyzed fusion schemes averaged on the four probe sets in the Biosecure database

Level	Type	Rule	Accuracy
score	fixed	sum	99.36%
score	fixed	min	80.56%
score	fixed	max	97.44%
score	fixed	median	97.86%
score	fixed	product	95.73%
rank	fixed	borda count	92.31%
rank	fixed	highest rank	81.62%
decision	fixed	majority voting	86.11%
score	trained	weighted sum	93.16%
hybrid	trained	predictor-based majority voting	97.22%
hybrid	trained	predictor-based borda count	92.52%
hybrid	trained	predictor-based sequential	96.58%

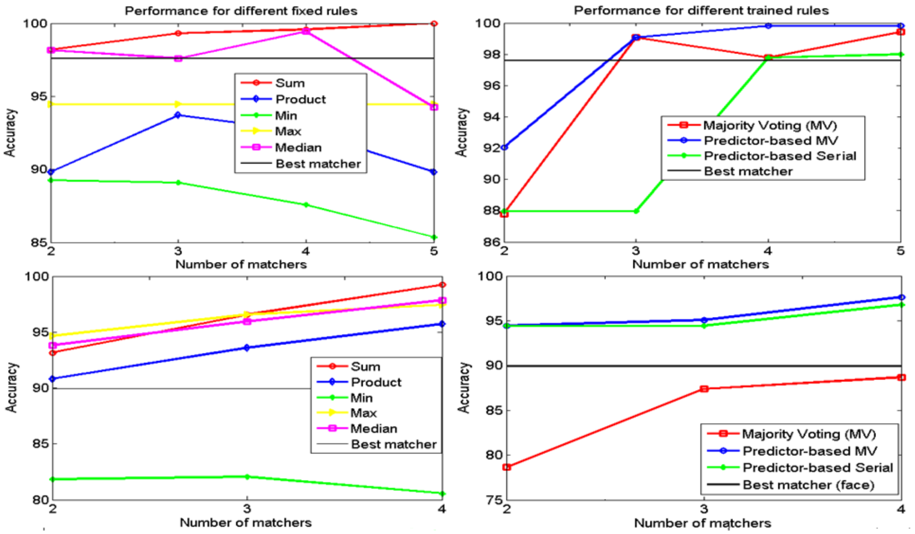


Fig. 1. Performance of fixed and trained fusion rules by varying the number of modality matchers: a) plots on the top show our results on WVU database which are obtained by averaging on the three probes, b) plots on the bottom show our results on Biosecure database which are obtained by averaging on the four probes

2 up to 5, while on Biosecure it ranges from 2 up to 4, and any considered subset was composed by the modality matchers having the best unimodal performance.

Regarding the performance of the fixed rules, our experiments showed that adding modalities to the fusion does not always imply increasing the multimodal performance. However we observed that, for the score sum adding modalities aims to increase the multimodal performance on both databases. On WVU database, the accuracy achieves 98.33% when two matchers are combined and increases to 100% when all the available matchers are combined, while on Biosecure database the accuracy achieves 93.33% when two matchers are combined and increases to 99.33% when all the available matchers are combined. Further, the max rule presents the same performance on the considered subsets of experts on WVU database, while on Biosecure the highest accuracy is achieved by using all the three matchers (97.44%).

In particular, fusion schemes at hybrid rank-score level always are able to improve the performance of the fusion schemes at rank level. The highest multimodal performance is obtained on WVU using trained rules.

4 Conclusions and Future Work

In this paper, we investigated how the benefits of the fusion change in identification scenario by varying the set of the fused modalities, the adopted fusion scheme and the performance of the individual matchers. The experimental comparison was carried out on two multimodal databases. Our experiments showed that, adding modalities to the fusion does not always imply increasing the multimodal performance. Our experiments showed also that, the improvement achievable by employing a multimodal solution depends on the adopted rules and on the data. Future research concerning this topic may regard a comparison of the considered schemes in presence of spoof attacks in order to study their security.

References

1. Ross, A., Jain, A.: Information fusion in biometrics. Pattern Recognition Letters 24, 2115–2125 (2003)
2. Poh, N., Kittler, J.: Multimodal information fusion. Multimodal Signal Processing: Theory And Applications For Human-Computer Interaction (2009)
3. Sanchez, U., Kittler, J.: Fusion of talking face biometric modalities for personal identity verification. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 5, pp. V-V (2006)
4. Grother, P., Tabassi, E.: Performance of biometric quality measures. IEEE Transaction On Pattern Analysis and Machine Intelligence 29(4), 531–543 (2007)
5. Kittler, J., Poh, N., Fatukasi, O., Messer, K., Kryszczuk, K., Richiardi, J., Drygajlo, A.: Quality dependent fusion of intramodal and multimodal biometric experts. In: SPIE Biometric Technology for Human Identification IV, vol. 6539 (2007)
6. Roli, F., Kittler, J., Fumera, G., Muntoni, D.: An experimental comparison of classifier fusion rules for multimodal personal identity verification systems. In: Roli, F., Kittler, J. (eds.) MCS 2002. LNCS, vol. 2364, pp. 325–336. Springer, Heidelberg (2002)

7. Nandakumar, K., Chen, Y., Dass, S., Jain, A.: Likelihood ratio-based biometric score fusion. IEEE Transaction on Pattern Analysis and Machine Intelligence 30(2), 342–347 (2008)
8. Ross, A., Nandakumar, K., Jain, A.: Handbook of MultiBiometrics. Springer, Heidelberg (2006)
9. Jain, A., Nandakumar, K., Ross, A.: Score normalization in multimodal bio-metric systems. Pattern Recognition 38(12), 2270–2285 (2005)
10. Poh, N.: Multi-System Biometric Authentication: Optimal Fusion And User-Specific Information. Ecole Polytechnique Fédéral de Lausanne (2006)
11. Vatsa, M., Singh, R., Noore, A., Ross, A.: On the dynamic selection of biomet-ric fusion algorithms. IEEE Transaction on Information Forensics and Security 5(3), 470–479 (2010)
12. Abaza, A., Ross, A.: Quality-based rank level fusion in biometrics. In: Third IEEE International Conference on Biometrics: Theory Applications and Systems (September 2009)
13. Marasco, E., Ross, A., Sansone, C.: Predicting identification errors in a multibiometric system based on ranks and scores. In: Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (September 2010)
14. Kuncheva, L.I.: Combining Pattern Classifiers Method and Algorithms. Wiley, Chichester (2004)
15. Marcialis, G.L., Roli, F.: Serial fusion of fingerprint and face matchers. In: Haindl, M., Kittler, J., Roli, F. (eds.) MCS 2007. LNCS, vol. 4472, pp. 151–160. Springer, Heidelberg (2007)
16. Crihalmeanu, S., Ross, A., Schuckers, S., Hornak, L.: A protocol for multibiometric data acquisition, storage and dissemination. Technical Report. West Virginia University (2007)
17. Poh, N., Bourlai, T., Kittler, J.: A multimodal biometric test bed for quality-dependent, cost-sensitive and client-specific score-level fusion algorithms. Pattern Recognition 43, 1094–1105 (2010)

Using Geometric Constraints to Solve the Point Correspondence Problem in Fringe Projection Based 3D Measuring Systems

Christian Bräuer-Burchardt, Christoph Munkelt, Matthias Heinze,
Peter Kühmstedt, and Gunther Notni

Fraunhofer Institute Applied Optics and Precision Engineering, Jena, Germany
{christian.braeuer-burchardt,christoph.munkelt,matthias.heinze,
peter.kuehmstedt,gunther.notni}@iof.fraunhofer.de

Abstract. A new method for fringe projection based 3D stereo scanners is introduced which realizes point correspondence finding and subsequent unwrapping of phase images without binary codes. The novelty of the method is the combination of geometric constraints between the three optical sensor components together with the estimated measurement accuracy of the system in order to achieve unique point correspondences. Considerable fringe code reduction obtained by use of geometric constraints and Gray code omission leads to a speed-up of the image sequence acquisition time. This opens the possibility to design moving sensors and to measure moving objects.

Keywords: fringe projection, phase unwrapping, optical measurement systems, measuring accuracy, epipolar geometry.

1 Introduction

Contactless metrology systems based on fringe projection technique are increasingly used for industrial, technical, and medical applications. Although flexibility, measuring accuracy, measurement data volume, and fields of application always increase, processing time should be reduced. Projection of structured light is used in many measuring systems for 3D surface acquisition. Recently considerable progress has been achieved in the field of contactless optical 3D measurements.

An extensive survey over coded structured light techniques to solve the correspondence problem which is the basis for 3D surface reconstruction is given by Battle et al. [1]. Coded structured light does not only include fringes but also other patterns like barcodes or stochastic patterns. In our work, however, we use exclusively fringe patterns. Methods using projection of sinusoidal fringe patterns have to solve the problem of phase unwrapping. This can be realized e.g. by the use of multiple spatial frequencies [2], temporal phase unwrapping methods [3], or use of Gray code sequences [4]. Due to its unambiguousness, use of Gray code leads to robust results. However, longer image sequences must be recorded, which limits the possible applications. We aim for significant reduction of the number of images of a structured light pattern sequence by omission of the Gray code and to speed-up the image recording time of

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 265–274, 2011.

fringe projection 3D measurement systems. This leads to non-static applications and therefore real-time applications or moving objects measurement become possible.

Zhang and Yau suggest a real-time coordinate measurement [5] where phase unwrapping is realized by determination and tracking of a marker. An interesting method for phase unwrapping using at least two cameras is presented by Ishiyama et al. [6]. There the number of possible correspondences is drastically reduced by back-propagation of the correspondence candidates into the image of the second camera. Ishiyama gives another suggestion for 3D measurement using a one projector one camera fringe projection system [7] using the invariance of cross-ratio of perspective projection. Young et al. [8] suggest the use of the limitation of the measuring volume in order to reduce the search area for corresponding points on the epipolar line to segments achieving a reduction of the projected binary code by careful placement of additional cameras (or additional measuring positions). Li et al. [9] use this approach in combination with the multi-frequency technique in order to realize real-time 3D measurements.

In our recent work we already achieved code reduction by using epipolar constraint and measuring volume restriction (see [10, 11]).

In this paper, an algorithm is presented which realizes point correspondence finding and phase unwrapping of single frequency fringe patterns without use of the Gray code at all which reduces the length of the projected fringe pattern sequences considerably. This is useful for applications where a short recording time of the image sequences is necessary as for example reconstruction of living objects (e.g. face recognition) and other measurement tasks which require real-time processing.

2 Model and Approach

2.1 Situation and Model

A number of fringe projection systems for 3D surface determination were developed at our institute [12, 13]. All of them base on the projection and observation of two orthogonal fringe sequences consisting of Gray code (5 to 7 images) and sinusoidal fringe patterns (between 3 and 16 images). From these sequences phase values are determined using a 4-, 8-, or 16-phase algorithm [13]. Phase values are used together with the extrinsic and intrinsic parameters of the optical components to determine the 3D point coordinates of the reconstructed object surface. Using epipolar constraint in order to find point correspondences is a typical task in phototogrammetry [14]. It reduces the task to a correspondence problem between two one-dimensional vectors.

Using fringe projection the difficult and uncertain task of correspondence finding (CFT) based on intensity correlation may become unambiguous using a robust phase unwrapping method, e.g. Gray code. However, 3D measurement systems based on fringe projection provide more geometric information which can be used for solving the CFT without using Gray code.

Let us consider using a sensor consisting of two cameras C_1 and C_2 and one projector P in a fix geometric arrangement (see fig. 1). Intrinsic and extrinsic parameters of all three components are known achieved either by self-calibration using current measurement data [15] or by a-priori calibration.

Each point *p* in the camera C_1 maps a 3D point *M* and corresponds to a straight line *g* in the image of camera C_2. If the measurement volume *mv* is restricted, the corresponding area of the line *g* is reduced to a segment *s* (see fig. 2). Point *q* corresponding to point *q* is expected to lie on *s*. A further restriction of the corresponding point candidates is achieved by the phase value obtained from the projected fringe pattern sequence [13, 15]. Since we want to omit the Gray code sequence which would make the CFT unique we use the rough phase values occurring periodically with a period length depending on the projected pattern.

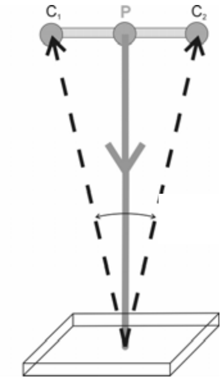


Fig. 1. Sensor arrangement

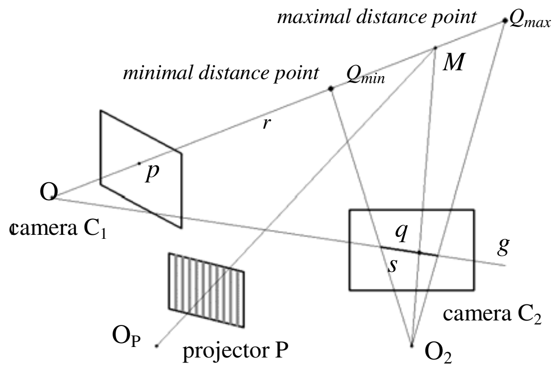


Fig. 2. Epipolar geometry with segment *s*

We typically use a projection display with an active area of 1024 x 1024 pixels, a period length of 16 pixels according to 64 fringes in maximum over the measurement volume. The number of projected fringes and the measurement volume size determine the number of expected periods mapped on the considered line segment *s*.

We assume the calibration including lens distortion determination (see [14, 16]) being completed and the parameters projection center O = (X, Y, Z), orientation angles ϕ, ω, and κ resulting in rotation matrix R, camera constant and principle point coordinates, and distortion description D of all components being known.

Ray *r* is defined in the 3D space by the projection center, the orientation of camera C_1, and point *p*. The predefined measurement volume *mv* is assumed to be a convex polyhedron, say a rectangular solid in a meaningful position of the object space. The two intersection points of *r* and *mv* result in two 3D points Q_{min} and Q_{max} as shown in fig. 2. The image points q_{min} and q_{max} in the image of camera C_2 are the undistorted segment endpoints of *s*. Hence applying distortion operator D_2 (of camera C_2) leads to the points $q'_{min} = D_2(q_{min})$ and $q'_{max} = D_2(q_{max})$ of segment *s'*. For simplification in the following we assume no distortion being present (obtained e.g. by perfect correction, see [14, 16]) and hence $s = s'$.

In order to obtain q_{min} and q_{max} first the ray *r* can be determined using O_1 and *p*. Then we get Q_{min} and Q_{max} from *r* and *mv*. Finally, we get q_{min} and q_{max} by applying the collinearity equations [14] using calibration data. For illustration see fig. 3.

The recording of the projected fringe sequence and calculation of the phase values leads to the distribution of the phase values on the epipolar segment *s* (see fig. 2). If no occlusions are present, each projected fringe period provides one candidate to be the corresponding point q to p.

Only in case of a very small measurement volume in direction of the optical axis or a very wide fringe width the number of mapped fringe periods on the epipolar segment is one and the point correspondence is unique. In general case a number of possible candidates q_i having the same rough phase value ϕ_p correspond to p leading to n possible 3D point candidates Q_i for the actual point Q.

However, the number of possible candidates corresponding to the number N of projected fringe periods can be drastically reduced using further geometric properties of the sensor. Let us consider a back-projection of all 3D point candidates q_i onto the projectors image plane leading to phase values ϕ_i. If the point correspondence is true, the origin ϕ_i in the projector image plane of the point q_i should correspond to the observed phase value ϕ_p: $\phi_i - \phi_p = 0$. If candidate q_i is false, the measured phase value ϕ_i may differ from the expected one ϕ_p. If the amount $|\delta_i|$of the difference $\delta_i = \phi_i - \phi_p$ is above a threshold *thr* candidate q_i will be rejected (see fig. 3).

Rejection of false candidates is, unfortunately, not complete. However, the expected number of remaining false candidates can be estimated using statistical properties. Let us consider the general case of an arbitrarily shaped measuring object. The general measuring accuracy of the system using wrapped phases leads to a typical distribution of the ϕ_d values of the correct correspondences. Here the mechanic and thermic stability of the sensor, the quality of distortion correction, and other conditions influence the measuring accuracy. Assume we have a certain accuracy of the back-propagation phase measurement error, characterized by the standard deviation *sd* of ϕ_d. Phase values should be in the interval $[0, 2\pi]$.

Assuming a normal distribution of the δ values of correct correspondences, 99.9% have an absolute difference value of $|\delta_q| < 3sd$. Choosing as threshold $thr = 3sd$ and assuming a uniform distribution of the phase values of the false candidates. This means, that every false candidate has a probability of $prob = 3sd/\pi$ to be below the threshold *thr* of rejection. Multiplied with the average number n of periods on one epipolar segment s minus one it gives the expected number *efc* of additional false candidates per true candidate:

$$efc = 3sd\left(n-1\right)/\pi \,. \tag{1}$$

Using equation (1) we can initially estimate the expected number of remaining false candidates to be rejected in a subsequent step. The amount of *efc* indicates the probability of the algorithm described in the following to be successful. The lower the amount of *efc* the better the correspondence finding algorithm will work.

2.2 Approach for Obtaining Unique Correspondences

In this section the approach for the procedure of resolving ambiguities and obtaining unique point correspondences is described.

In a first step all candidates concerning the δ- criterion ($|\delta_i| < thr$) are collected. Each point in the C_1 camera may have zero, one, or more corresponding candidates. In the first case there may be no correspondence because of hidden scene parts, image disturbances, or exceedance of the difference threshold *thr*. These points have no relevance except leading to incomplete regions in the resulting 3D point cloud.

Those pairs of corresponding point candidates without additional candidate are true correspondence with high probability. However, there may be also false correspondences, namely in case of absence of the true corresponding point for point p_1.

In the third group of multiple corresponding candidates for one point p_1 the true correspondence is present in the set of candidates with high probability, except in the case described before.

Let an initial set of points in the image of camera C_1 and a set of all initially found (still multiple) corresponding points in the image of C_2 be given. A number of unique correspondences can be identified due to processing as explained above. To maximize the number of unique correspondences, the following steps are performed:

1. Selection of reference correspondences (assumed to be correct)
2. Selection of correspondences out of the set of points with multiple corresponding candidates
3. Decision whether uncertain correspondences are true or false
4. Rejection of correspondences by geometric contradictions (section 2.2.3)

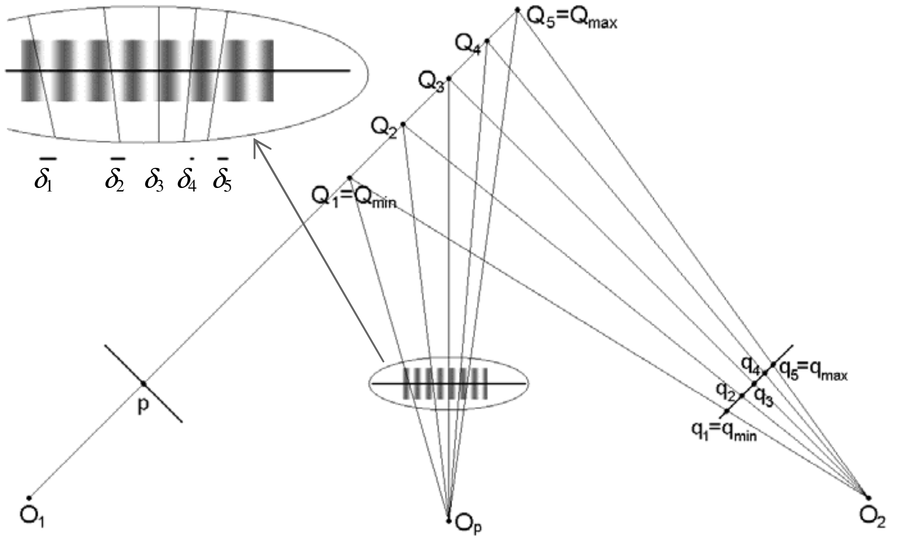


Fig. 3. Point p and five candidates q_i: three candidates (q_1, q_2, q_5) are rejected because of too big difference values δ_i, two candidates remain: correct one (q_3) and false one (q_4) – view from above. The phase values on projector image plane are represented by sinusoidal fringe pattern, the length of horizontal bars above the 'δ_i' represents the amount of error δ_i. Error δ_3 is assumed to be zero.

2.2.1 Selection of Reference Correspondences

For the selection of reference correspondences (step 1) no fix algorithm may be given here, because the conditions of the measurement are different. However, some aspects should be noticed. First, an analysis of the expected number of false candidates according to equation (1) should be performed taking into account the expected accuracy of the projector phase and expansion of measurement volume. Additionally the expected number of unfound correspondences (occlusions, shadows, etc.) may be roughly estimated. The percentage of unique correspondences *puc* yields from equation (1) and is about $puc \approx 100 \cdot (1 - efc)$ in the case of *efc* considerable smaller than one. The following properties are possible criteria for reference points.

- Exactly one corresponding candidate exists
- Neighbouring correspondences are similar (e.g. small difference in coordinates) and unique, too
- No violation of the monotony of the unwrapped phase and the corresponding coordinate along the epipiolar segment

2.2.2 Selection of Correspondences of the Set of Points with Multiple Corresponding Candidates and Evaluation of Uncertain Points

After the reference points are identified all remaining points are addressed. There are still uncertain candidates having exactly one correspondence and points with multiple corresponding candidates. Let the current candidate corresponding to point *p* in the C_2 image be $c(p) = q$. Consider a small surrounding of the corresponding epipolar segments s_1 and s_2 with *p* and *q* as center-points, respectively. The phase values at positions *p* and *q* are denoted by $\phi(p)$ and $\phi(q)$. Search for reference correspondences around *p* and *q* (see fig. 4) with maximal phase difference of 2π with coordinates k_i, g_i, $c(k_i)$, and $c(g_i)$, respectively. If the correct corresponding point to *p* is *q*, it must hold for phase values ϕ of all reference correspondence coordinates $c(k_i)$ and $c(g_i)$:

$$\begin{gathered}\varphi(k_i) < \varphi(p) \wedge \varphi(p) - \varphi(k_i) < 2\pi \\ \varphi(g_i) > \varphi(p) \wedge \varphi(g_i) - \varphi(p) < 2\pi \,. \\ \varphi(c(k_i)) < \varphi(q) \forall k_i \\ \varphi(c(g_i)) > \varphi(q) \forall g_i\end{gathered} \tag{2}$$

If this holds for just one candidate, it must not be valid for all the other possible candidates. However, there may be contradictions for some candidates or no unique decision possible. In these cases correspondence must be fixed in a following step.

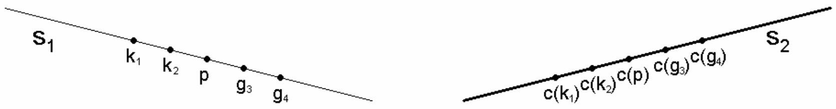


Fig. 4. Evaluation of uncertain point correspondence candidates *c*(*p*) according to (2)

2.2.3 Rejection of Found Correspondences by Contradictions and Additional Decisions

All found correspondences may be tested using the feature of the monotony of the coordinates (only one coordinate is monotone!) and of the unwrapped phase values along the epipolar lines over the considered image areas. Contradictions may be found and monotony may be enforced. However, this should not be explicitly described here.

If the corresponding pairs of reference are not well distributed, which is, unfortunately the typical case, a number of correspondences remains open (more than one candidate is possible) or uncertain (current correspondence may be false). Decisions must be realized using other criteria. Here the regions are extended also perpendicular to the direction of the epipolar lines. The complete correspondence finding will be obtained by iterative application of the described procedure.

2.3 Realization of the Method

The method described so far was implemented in a test environment using measurement data from a measuring device at our institute. The results are given in the next section. Here the necessary preparation tasks are described.

First, initial analyzing experiments were performed including estimation of the back-propagation phase accuracy in order to set the threshold *thr*.

The implementation of the described algorithm was done by replacing the part of finding corresponding points in the current software package of our measuring system. As we assume well-known calibration values including the distortion function, the epipolar segments do not depend on the current measurement. Hence the corresponding segments to all image points p of C_1 can be determined before the measurement. The endpoints of the segments are stored in suitable data files. Point correspondence finding is performed for each point p in the image of camera C_1. The value of the rough phase ϕ_p is determined and the same phase values are searched for on segment s. All candidates q_i with a small difference value ($|\phi_i)| < thr$) become correspondence candidates. The selection of the true candidate is performed as described in the previous section.

3 Experiments and Results

The developed algorithm was implemented and tested using datasets obtained by the measuring device “kolibri Flex mini” [16]. This measuring system is a table-top system for measuring objects up to size of 80 mm (diameter) x 25 mm (height).

Rough phase data files were generated using unwrapped phase values obtained by a 16-phase algorithm (see [2]). The mean number n of fringe periods per segment was $n = 6$, and the mean standard deviation sd of the back-propagation phase measurement error was $sd = 0.041$ leading to an expected mean number of remaining false candidates of $efc = 0.20$ per correspondence.

In order to evaluate the developed method some experiments were performed. The goal was to get results in completeness, correctness, and accuracy. A number of different measuring objects were selected including a toy elephant, a prism, a single plaster tooth, a machine tool (see fig. 5), a set of teeth made of plaster, a plaster nose, and a plane surface. In order to compare the results to those of applying an algorithm including Gray code sequences a reference dataset (RD) was produced for every measurement including point correspondence data and 3D surface points.

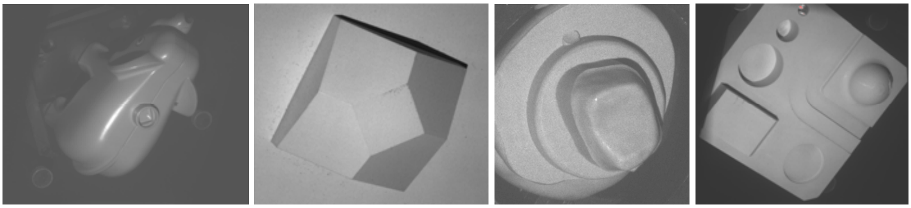


Fig. 5. Selected measuring objects: elephant, prism, plaster tooth, machine tool

The resulting point correspondences applying the new method were classified concerning the following characteristic features:

- Completeness *com* (percentage of the correctly found point correspondences concerning the number of point correspondences in RD)
- Completeness error *miss* (percentage of unfound correspondences)
- Correspondence error *fpc* (percentage number of false positive point correspondences)

Because of the possibility of finding all correct point correspondences but having additionally some false positive point correspondences, the sum of $com + miss + fpc$ may exceed 100%. Additionally, found point correspondences were used to produce 3D measurement data. These datasets were used to compare the measurement results to the reference results. The results for the characteristic features *com*, *miss*, and *fpc* are presented in table 1. The higher error of the nose object is among others (hidden parts) due to extended height (about 45 mm) of the object leading to higher *efc* value.

Sequence length was significantly reduced according to the method used so far (see table 2). Some reconstructed object surfaces obtained by processing of the 3D point clouds are shown in fig. 6.

Table 1. Completeness Results

Object	*com* in %	*miss* in %	*fpc* in %	number of points
Plane	100.00	0.00	0.00	6384
Prism	99.96	0.03	0.02	19307
Machine tool	99.99	0.01	0.03	28568
Set of teeth	99.99	0.01	0.03	70249
Plaster tooth	100.00	0.00	0.13	6769
Nose	99.92	0.06	2.02	22709
Elephant	99.99	0.01	0.00	11412

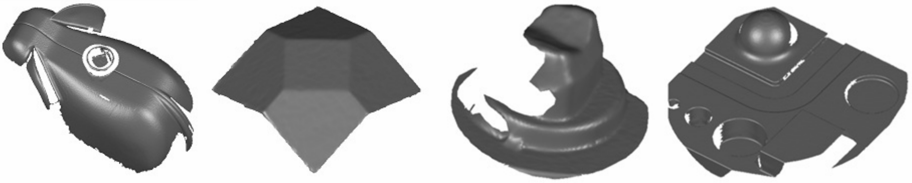


Fig. 6. Results (3D surfaces) of selected measuring objects

Table 2. Sequence length (number of images) according to several phase algorithms [2] (GC = 7 image Gray code), PA achieves reduction to 70%, 54%, and 37% concerning EGC

Algorithm	16-phase	8-phase	4-phase
Conventional, two directions, full GC (CGC)	46	30	22
Full GC with epipolar lines (EGC)	23	15	11
Proposed algorithm without GC (PA)	16	8	4

4 Summary, Discussion, and Outlook

We introduced a new method for fringe projection based 3D stereo scanners is introduced which realizes point correspondence finding and subsequent unwrapping of phase images without binary codes. It works robust if certain geometric properties hold. The geometric constraints of all optical sensor components were used in combination of the expected measuring accuracy of the sensor. A considerable reduction of the projected fringe code can be achieved by omission of the entire Gray code sequence resulting in a faster image sequence recording. This is important for high speed applications, e.g. in handheld 3D measurement systems.

It could be shown experimentally, that a completeness of 99.9% and an error rate of about 0.1 % in the case of *efc* = 0.2 and 2.0 % at *efc* = 0.5 can be achieved depending also on the object properties. This is sufficient for practical applications because the 3D points resulting from false positive point correspondences may be eliminated in the 3D space by suitable outlier detection algorithms. However, we assume a worse performance of the algorithm at values of *efc* greater than about 1.0 corresponding to an extended measuring volume depth. We recommend applying the method in the case of *efc* smaller than about 0.5. Alternatively, fringe period length can be extended or, if possible, back-propagation phase error must be reduced by some algorithmic or hardware improvement.

The image recording time can be considerably reduced compared to conventional image recording using Gray code and providing the same accuracy of the measurement. This leads to significant saving of measurement time. On the other hand, the point correspondence finding algorithm as described needs more computational effort than using Gray code sequences. Because the algorithm needs to be applied iteratively, the computation time may considerably vary. Hence an estimation of additional computation time is difficult.

However, computation time in average and worst case, respectively, will be analysed in a next step. Furthermore, calculation time of the algorithm has to be optimized. Additionally, future work should also include implementation of the algorithm into several measuring systems for 3D surface measurement using fringe projection and performing experiments using other sensor systems in order to evaluate the algorithm.

References

1. Battle, J., Mouaddib, E., Salvi, J.: Recent progress in coded structured light as a technique to solve the correspondence problem: a survey. PR 31(7), 963–982 (1998)
2. Li, E.B., Peng, X., Xi, J., Chicaro, J.F., Yao, J.Q., Zhang, D.W.: Multi-frequency and multiple phase-shift sinusoidal fringe projection for 3D profilometry. Optics Express 13, 1561–1569 (2005)

3. Zhang, H., Lalor, M.J., Burton, D.R.: Spatiotemporal phase unwrapping for the measurement of discontinuous objects in dynamic fringe-projection phase-shifting profilometry. Applied Optics 38, 3534–3541 (1999)
4. Sansoni, G., Carocci, M., Rodella, R.: Three-dimensional vision based on a combination of Gray-code and phase-shift light projection: Analysis and compensation of the systematic errors. Applied Optics 38(31), 6565–6573 (1999)
5. Zhang, S., Yau, S.T.: High-resolution, real-time 3D absolute coordinate measurement based on a phase-shifting method. Optics Express 14(7), 2644–2649 (2006)
6. Ishiyama, R., Sakamotom, S., Tajima, J., Okatani, T., Deguchi, K.: Absolute phase measurements using geometric constraints between multiple cameras and projectors. Applied Optics 46(17), 3528–3538 (2007)
7. Ishiyama, R., Okatani, T., Deguchi, K.: Precise 3-d measurement using uncalibrated pattern projection. In: Proc. IEEE Int. Conf. on Image Proc., vol. 1, pp. 225–228 (2007)
8. Young, M., Beeson, E., Davis, J., Rusinkiewicz, S., Ramamoorthi, R.: Viewpoint-coded structured light. In: Proc. CVPR, pp. 1–8 (2007)
9. Li, Z., Shi, Y., Wang, C.: Real-time complex object 3D measurement. In: Proc. ICCMS, pp. 191–194 (2009)
10. Bräuer-Burchardt, C., Munkelt, C., Heinze, M., Kühmstedt, P., Notni, G.: Phase unwrapping in fringe projection systems using epipolar geometry. In: Blanc-Talon, J., Bourennane, S., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2008. LNCS, vol. 5259, pp. 422–432. Springer, Heidelberg (2008)
11. Bräuer-Burchardt, C., Munkelt, C., Heinze, M., Kühmstedt, P., Notni, G.: Fringe code reduction for 3D measurement systems using epipolar geometry. In: Proc. PCVIA, ISPRS, vol. XXXVIII, Part 3A, pp. 192–197 (2010)
12. Kühmstedt, P., Heinze, M., Himmelreich, M., Bräuer-Burchardt, C., Notni, G.: Optical 3D sensor for large objects in industrial application. In: Proc. SPIE, vol. 5856, pp. 118–127 (2005)
13. Kühmstedt, P., Munkelt, C., Heinze, M., Bräuer-Burchardt, C., Notni, G.: 3D shape measurement with phase correlation based fringe projection. In: Proc. SPIE, vol. 6616, 66160B (2007)
14. Luhmann, T., Robson, S., Kyle, S., Harley, I.: Close range photogrammetry. Wiley Whittles Publishing, Chichester (2006)
15. Schreiber, W., Notni, G.: Theory and arrangements of self-calibrating whole-body three-dimensional measurement systems using fringe projection technique. Opt. Eng. 39, 159–169 (2000)
16. Bräuer-Burchardt, C., Heinze, M., Munkelt, C., Kühmstedt, P., Notni, G.: Distance dependent lens distortion variation in 3D measuring systems using fringe projection. In: Proc. 17th BMVC, pp. 327–336 (2006)

Retrospective Illumination Correction of Greyscale Historical Aerial Photos

Anders Hast* and Andrea Marchetti

Consiglio Nazionale delle Ricerche (CNR), Institute of Informatics and Telematics (IIT)
Via Moruzzi, 1 – CNR, Research Area
Pisa, Italy
{Anders.Hast,Andrea.Marchetti}@iit.cnr.it

Abstract. Illumination correction is a method aiming at removing the influence of light from the environment and other distorting factors in the image capture process. A novel algorithm based on luminance mapping is proposed that both removes the low frequency variations in intensity as well as increases the contrast in low contrast areas. Moreover, it avoids the common problems with homomorphic filters. This algorithm is being applied on historical aerial photos with good results.

Keywords: Illumination Correction, Luminance Mapping, Image Stitching, Image Mosaicing, Vignetting.

1 Introduction

In many disciplines dealing with images the problem of varying illumination of the scene or object must be faced. Hence both the problem and the remedy have got different names depending on discipline. Vignetting [24, 26] occurs due to different mechanisms [7], which causes a brightness falloff away from the image centre and this is prevalent in photography. Non uniform illumination or even dirty lenses and dust [23] causes problems in Microscopy [11]. Such procedure is often called illumination or shading correction and in this paper we will refer to it as *Illumination Correction*. In Magnetic Resonance Imaging (MRI) the varying shade is known as RF-inhomogeneity or bias [1,2]. In Face Recognition the varying illumination is a challenging problem [25, 27] and the method is, besides illumination correction also called illumination normalization. The illumination itself is often referred to as background light. In Mammography contrast-limited adaptive histogram equalization (CLAHE) has been used [17].

In aerial and satellite photos the physical lighting also affects the relief presentation [19, 20]. However it is also a problem for image stitching and mosaicing of panoramas where the focus often is put on making the transition from one image to the other as smooth as possible [4, 12, 15].

* This work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 275–284, 2011.

In this paper we will focus on a method that we developed for *retrospective illumination correction of greyscale historical aerial photos* that will not only minimize the variance in illumination over the image but can also be used to increase contrast in low contrast areas in the image.

1.1 Main Methods

As mentioned in the introduction the same problem of uneven illumination occurs in many fields and the referenced papers suggest different methods specialized for the application in question. However there are some fundamental aspects that will be shortly discussed here.

Theoretically, the intensity in the image can be divided into two components, the illumination of the object and the amount of light reflected by the object [14]. The homomorphic filter [13] aims at removing the illumination, characterized by low spatial variations, from the image and keeping reflectance, characterized by high frequency changes. In terms of the *illumination L(x,y)* and *reflectance R(x,y)*, the *intensity H(x,y)* of a pixel can be modelled as:

$$H(x,y) = L(x,y)R(x,y). \tag{1}$$

The homomorphic filter makes use of the Fourier transform to remove the *intensity* and hence it is necessary to convert the multiplication to addition, which can be done by taking the log of the functions, as the Fourier transform only can be used when the noise is an additive term. This approach has some disadvantages, for instance the necessary image size padding when the image size is not a power of 2, which can lead to distortions in the boundary regions [5]. Another problem with this approach is that it will merely increase contrast in low contrast regions while the main low frequencies remains. A better way to look at the problem [27] is to model the lighting change as a local affine transformation (AT) of the pixel value:

$$H(x,y) = A(x,y)h(x,y) + B(x,y). \tag{2}$$

where *h(x,y)* is the original image and *H(x,y)* is the captured image as a result of the local lighting, which thus have both and multiplicative and additive effect on the result. Zhu et al [27] removes the low spatial variations caused by *B(x,y)* using a multi resolution low pass filter that estimates the background lighting. Leung et al [11] takes the same approach using a Gaussian filter:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}. \tag{3}$$

where σ defines the width of the distribution used to delimit the low frequencies. The larger the value of σ, the greater the smoothing effect. Others, for an example Yu [24] handle the background light using a reference picture acquired in a controlled environment with known lighting conditions and a white object. This is of course not always possible, especially not in retrospect.

The gain *A(x,y)* in eq. 2 is handled by a homomorphic filter and this will increase the contrast. The approach by Zhu gives better results but the main disadvantages with the homomorphic filtering remains. The resulting image computed by AT is:

$$\breve{h}(x,y) = \frac{\kappa}{\breve{A}(x,y)}\left(H(x,y) - \breve{B}(x,y)\right). \tag{4}$$

Where $\breve{A}(x,y)$, $\breve{B}(x,y)$ are the estimate of $A(x,y)$, $B(x,y)$ respectively and κ is a constant. Zhu et al modifies this equation further but the equation shown here is the basis. Yucong et al [25] have improved the AT by computing the estimated gain as a low pass filtered estimate:

$$\breve{A}(x,y) = \exp\left(\mathrm{LPF}\left(\ln\left(h(x,y) - \breve{B}(x,y)\right)\right)\right). \tag{5}$$

where LPF is a low pass filter. They elaborate further on his approach using histogram equalization and also combining with ICR [9] explained shortly in next subsection, however these are the basic ideas.

1.2 Other Approaches

Others have developed methods for finding the *vignetting function* from a single image [26] or from multiple images, with or without the response curve [7]. The ICR (Illumination Compensation based on Multiple Regression Model) [9] is aiming to find the plane that best fits the intensity distribution of the image using the multiple regression model. Then this plane is used to remove the illumination of face images. The Local Range Modification (LRM) [6] , finds the interpolated minimum and maximum pixel values in a neighbourhood (contextual region) and stretches them to the desired range via the equation:

$$Y(x,y) = \frac{C}{max - min}(X(x,y) - min). \tag{6}$$

where *C* is a constant. Similarly the contrast-limited adaptive histogram equalization method (CLAHE) [16], stretches the histogram in its contextual region. This improves the image in such way that the contrast is enhanced. Young et al [25] make use of homomorphic filters and Morphological filtering in order to correct images in microscopy. Here the morphological filter is used to obtain an estimate of the background illumination.

Wavelet based approaches for image and contrast enhancement are also popular [18]. Laine et al [10] use wavelets to enhance contrast in digital mammography. Yet others make use of a combination of methods as wavelets and the homomorphic filter [3,22].

1.3 Luminance Mapping

Colour correction can be used to both correct and filter colours in an image [21]. One important method is Luminance mapping [8]. The idea is to apply a linear map that matches the means and variances of the luminance distributions, i.e. the intensity. If *A(x,y)* is the luminance of a pixel in image *A*, then it is remapped using the distribution of image *B*:

$$A(x,y) = \frac{\sigma_B}{\sigma_A}(A(x,y) - \mu_A) + \mu_B. \tag{7}$$

where μ_A and μ_B are the mean luminances, and σ_A and σ_B are the deviations of the luminances, both taken with respect to the luminance distributions in A and B, respectively. The result, if applied an all three colour channels, is that image A will have similar looking colours as image B. The result is shown in Fig. 1 where the original Lena and Mandrill images are to the left and the luminance mapped images to the right where Lena has got the colours from the Mandrill and vice versa. This method has not got any obvious connection to illumination correction, however it is the basis for the proposed method discussed in the next section.

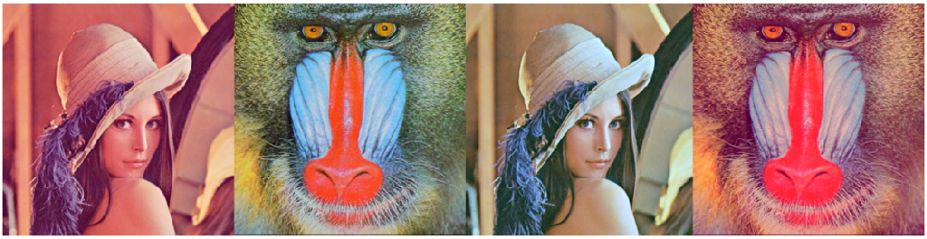


Fig. 1. The original Lena and Mandrill images to the left and the luminance mapped images to the right, where Lena has got the colours from the Mandrill and vice versa.

2 The Proposed Method

In the approach by Zhu et al the low frequencies are removed first by subtracting $\breve{B}(x,y)$, which is an estimation of the background lighting and then the contrast is increased by multiplication of $\kappa/\check{A}(x,y)$ (Note that they use a more elaborated version of this factor in their paper, including eq. 5). This implies that $\breve{B}(x,y)$ contains the low frequencies in the image. If eq. 7 is applied using, not an overall estimate of the mean but an estimate of the local mean µ it is possible to obtain a similar result. Let us take the image of a checkerboard in the left of Fig. 2 as an example, which has an obvious intensity variation. Looking at a horizontal cut section taken from the centre, depicted to the right, we can see the intensity variation from left to right. The *local mean* is the red curve that runs in the centre of the curve. If the *local mean* is translated to the *global mean*, i.e. pushing down the curve so that it becomes a line, then a big portion of the varying light is removed.

A *low pass filter* is used to remove an estimate of the illumination, which are the lowest frequencies in the image and thus an approximation of the *local mean* $\mu_A(x,y)$. This is done by using a Gaussian filter (eq. 3) modified in a way such as it also compensates for the borders. However the mask must be quite large, up to a size of the input image. Even with modern computers it becomes impractical for images of sizes up to 5000x6000 pixels, which we are dealing with in the case of historical aerial photos. This problem can be easily overcome by the fact that the estimate is a heavily blurred version of the input image and therefore we can compute $\tilde{\mu}_A(x,y)$ using a downscaled and blurred version $\widetilde{H}(x,y)$ of the image and then upscale it to $\check{\mu}_A(x,y)$ using a bilinear interpolation. The resulting blurred image will be very close to what the full scale Gaussian filter would produce and will not be detrimental for the final quality, however in Fig. 2-4 a scale factor of 1 was use, i.e. no downscaling at all.

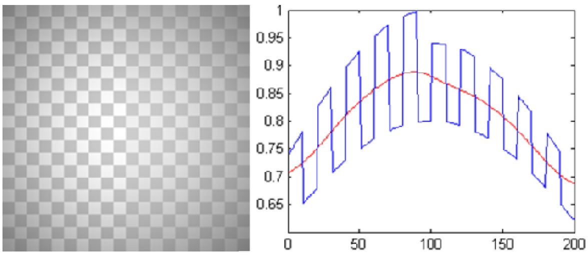


Fig. 2. The image to the left has an obvious shading artefact visible as an intensity fall off away from the centre. The diagram to the right shows a cut section in the middle in blue and the local mean in red.

Fig. 3 shows how the image in Fig.2 have been processed by the simple method of subtracting the local mean estimated by the blurred image and adding the global mean, in order to correct the illumination. It can be seen that the curve is corrected, however the contrast is low on the sides. In this case the image has been processed three times with varying size of the Gaussian mask (in sizes of the whole image: ¼, ½ and 1).

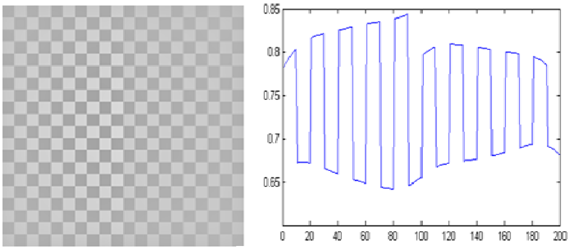


Fig. 3. After processing the illumination is corrected as also is shown to the left in the cut section. However, the contrast is not yet corrected.

In the next step the local deviation is computed as the difference between the local mean taken from the blurred image and the pixel values in a block region, whose size is equal to the Gaussian kernel. The block size was chosen the same as for the computation of the local mean for practical reasons, however there is nothing that prevents for testing other sizes. The local deviation $\tilde{\sigma}_A(x,y)$ is hence computed as the absolute value of the difference between $\tilde{\mu}_A(x,y)$ and the downscaled version $\widetilde{H}(x,y)$ of the original image (x,y) :

$$\tilde{\sigma}_A(x,y) = \left| \widetilde{H}(x,y) - \tilde{\mu}_A(x,y) \right| \tag{8}$$

The result $\tilde{\sigma}_A(x,y)$ is then subsequently blurred using the Gaussian mask and then upscaled to its full size $\breve{\sigma}_A(x,y)$. Therefore this algorithm is doing something similar as Yucong et al as it appears in eq. 5, however, with different mask sizes and without the logarithm and exponential function. Moreover it is used in a different and more simple way as it is plugged in to eq. 7 and we get:

$$\check{h}(x,y) = \frac{\sigma_B}{\check{\sigma}_A(\mathrm{x,y})}\big(H(x,y) - \check{\mu}_A(x,y)\big) + \mu_B. \tag{9}$$

This equation is used in three steps as before with the same sizes of the Gaussian kernel and the result is shown in Fig. 4. The proposed contrast enhancement using luminance mapping yields a straightened curve where the amplitude is almost the same over the whole curve and the shading artefact is removed from the image.

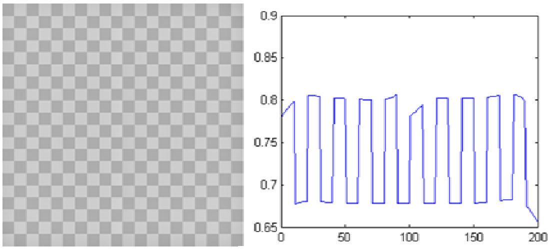
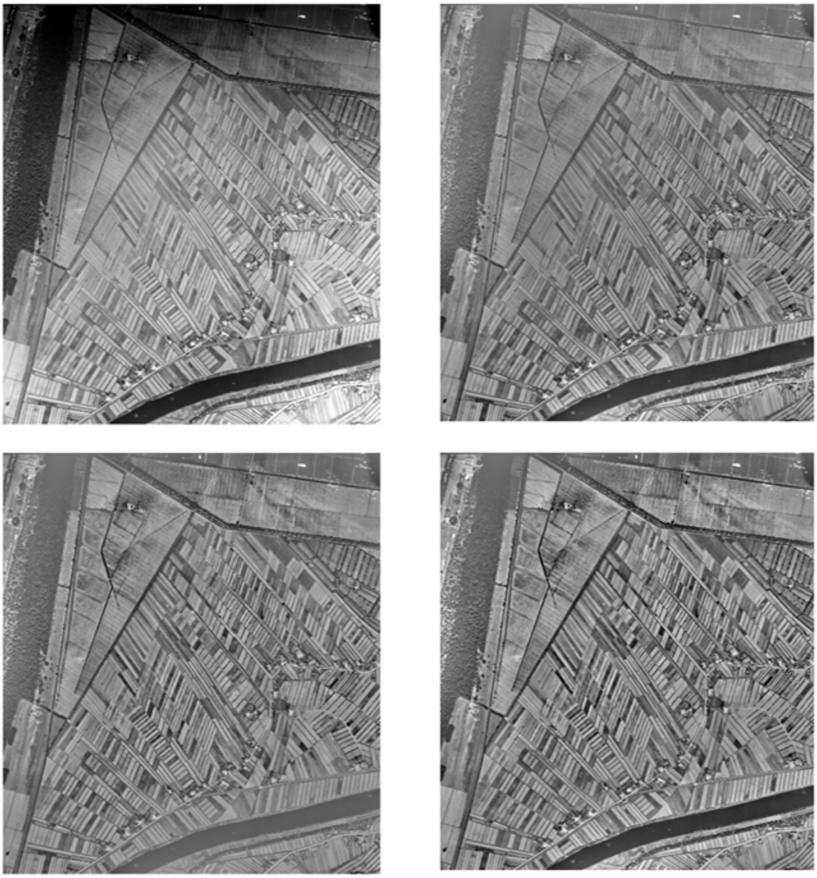


Fig. 4. After further processing the contrast is enhanced and the two halves become almost identical

3 Results

Let us examine some results of using the proposed method and look at some necessary changes to the algorithm. A historical photo showing the river Arno close to Pisa (Italy) taken during World War II is shown in Fig 5. On the top left is a cut out from the original photo (3800x3400 pixels). The illumination is obviously unevenly spread over the picture. On the top right is the process of pushing down the mean done three times, in the same manner as for the checkerboard in Fig.2-4. Since the light was either very low in the top or very high in the bottom, these regions will have quite low contrast. The change in the local mean of the image does not affect the contrast and therefore the contrast enhancement has been applied in the bottom row. The contrast is now higher in the top and in the bottom right of the image. However, in the same time as low contrast regions are getting increased contrast, the algorithm will is also decrease the already high contrast areas. Hence it is often necessary to avoid contrast enhancement in this areas. This can be easily done by not allowing $\check{\sigma}_A$ to be larger than σ_B, or not larger than some percentage of σ_B. In this manner we can adjust the image to obtain the desired result. The bottom left shows the result when high contrast regions are kept without lowering the contrast. A scale factor of 10 was used processing these images.

Moreover, the reason for repeating the process three times, starting with a Gaussian mask of size ¼ of the image and then doubling the size, is evident in this picture because, if we would have taken a smaller size, let us say $^1/_8$ of the image then the river Arno would become much brighter as the local mean would be closer to the true



© MiBAC-ICCD, Aerofototeca Nazionale, fondo RAF.

Fig. 5. A historical photo showing the river Arno close to Pisa taken during WW- II. On the top left is a cut out from the original photo (3400x3800 pixels), with unevenly spread illumination. On the top right is the mean corrected three times. Low contrast regions are apparent in the top and bottom of the image where the light was low or high. On the bottom row left, is the contrast increased, which also means that high contrast regions will have lower contrast. The bottom right shows the result when high contrast regions are kept without lowering the contrast.

value. This can be seen in Fig. 6, where the mask size is ¼ to the left and $^1/_8$ to the right. This is a vertical cross section of the top left image in Fig. 5. Clearly the red curve in the right picture is a better approximation of the blue curve, however it will lift up Arno, which is the dip in the curve to the right. Hence, it is necessary to use a larger mask size resulting in an image that is not fully corrected. Therefore, it is often necessary to repeat the process with a larger mask for each step to correct this behaviour. Experimentally we found that doubling the mask is generally a good approach for obtaining visually pleasing results.

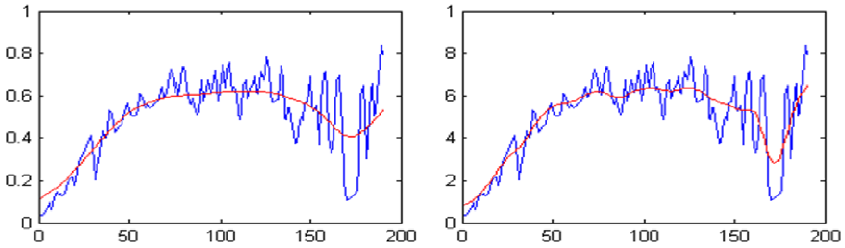
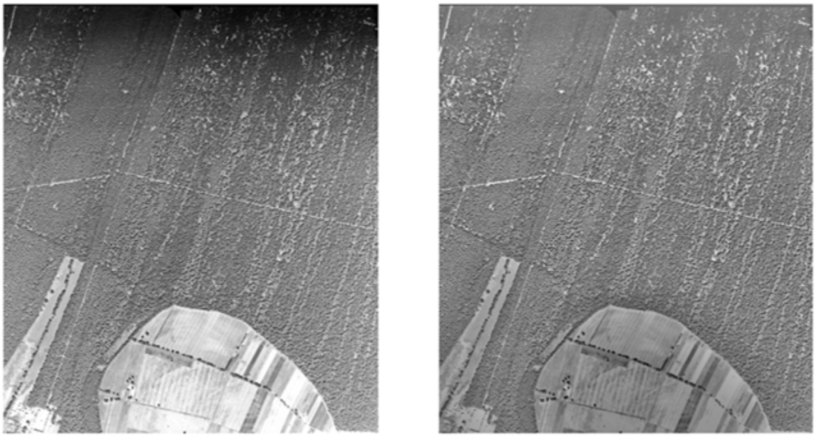


Fig. 6. A vertical cross section of the top left image in Fig. 5, where the mask size is ¼ to the left and $^1/_8$ to the right. A smaller mask gives a local mean (in red), which is a better approximation of the curve (in blue). However, it will also reduce the contrast in the image.

In fig. 7. There is another historical image taken during WW-II close to Pisa, showing an open field in the bottom with a forest surrounding it. The illumination is once again very uneven and the corrected result to the left has an even illumination.



© MiBAC-ICCD, Aerofototeca Nazionale, fondo RAF.

Fig. 7. A field close to Pisa during WW-II. To the left a cut out of the original image (3300x3800 pixels) and the corrected image to the right.

One can note a couple of things, for instance that the cut was done wrong in the sense that there is some parts of an arrow visible in the top that belongs to the border of the original image. Without the illumination correction it would have been hard for the human eye to see this. Moreover, the field becomes quite dark as the algorithm tend to set the even light over the whole picture and the field occupies a substantial part of it. Setting a smaller Gaussian mask helps a little but then the light will be less even in the image.

4 Conclusions

The proposed use of luminance mapping for illumination correction has not to our knowledge been done before. The other main contributions in this paper is that we propose a modification where both the local mean and deviation is obtained by a Gaussian filter, where both can be computed from downscaled versions of the original image and then upscaled by bilinear interpolation. This will decrease computation time as the Gaussian interpolations in computationally expensive, however it will not affect the result for moderate downscales. Moreover, it was shown that the contrast can easily be enhanced only where needed. The example images shows that the proposed approach of using a modified version of luminance mapping for illumination correction works very well for retrospective illumination correction of historical aerial photos.

References

1. Agus, O., Ozkan, M., Aydin, K.: Elimination of RF Inhomogeneity Effects in Segmentation. In: Proceedings of the 29th Annual International Conference of the IEEE EMBS, pp. 2081–2084 (2007)
2. Ardizzone, E., Pirrone, R., La Bua, S., Gambino, O.: Volumetric Bias Correction. In: Gagalowicz, A., Philips, W. (eds.) MIRAGE 2007. LNCS, vol. 4418, pp. 525–533. Springer, Heidelberg (2007)
3. Ashiba, H.I., Awadalla, K.H., El-Halfawy, S.M., Abd El-Samie, F.E.-S.: Homomorphic enhancement of infrared images using the additive wavelet transform. Progress In Electromagnetics Research C 1, 123–130 (2008)
4. Burt, P.J., Adelson, E.H.: A multiresolution spline with application to image mosaics. Journal ACM Transactions on Graphics (TOG) 2(4) (1983)
5. ERDAS Field Guide, p. 525 (2010), `http://www.erdas.com/Libraries/Tech_Docs/ERDAS_Field_Guide.sflb.ashx`
6. Fahnestock, J.D., Schowengerdt, R.A.: Spatially variant contrast enhancement using local range modification. Optical Engineering 22, 378–381 (1983)
7. Goldman, D.B., Chen, J.H.: Vignette and Exposure Calibration and Compensation. In: Proceedings of ICCV 2005, pp. 89–906 (2005)
8. Hertzmann, A., Jacobs, C.E., Oliver, N., Curless, B., Salesin, D.H.: Image Analogies. In: SIGGRAPH 2001 Conference Proceedings, pp. 327–340 (2001)
9. Ko, J., Kim, E.-J., Byun, H.: A simple illumination normalization algorithm for face recognition. In: Ishizuka, M., Sattar, A. (eds.) PRICAI 2002. LNCS (LNAI), vol. 2417, pp. 532–541. Springer, Heidelberg (2002)
10. Laine, A., Fan, J., Yang, W.: Wavelets for contrast enhancement of digital mammography. IEEE Engineering in Medicine and Biology Magazine 14(5), 536–550 (1995)
11. Leong, F.J.W.-M., Brady, M., O'D McGee, J.: Correction of uneven illumination (vignetting) in digital microscopy images. J. Clin. Pathol., 619–621 (2003)
12. Levin, A., Zomet, A., Peleg, S., Weiss, Y.: Seamless image stitching in the gradient domain. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3024, pp. 377–389. Springer, Heidelberg (2004)
13. Oppenheim, A., Schafer, R., Stockham Jr., T.: Nonlinear filtering of multiplied and convolved signals. IEEE Transactions on Audio and Electroacoustics 16(3), 437–466 (1968)

14. Pajares, G., Ruz, J.J., de la Cruz, J.M.: Performance analysis of homomorphic systems for image change detection. In: Marques, J.S., Pérez de la Blanca, N., Pina, P. (eds.) IbPRIA 2005. LNCS, vol. 3522, pp. 563–570. Springer, Heidelberg (2005)
15. Pérez, P., Gangnet, M., Blake, A.: Poisson Image Editing, Journal ACM Transactions on Graphics (TOG). In: Proceedings of ACM SIGGRAPH, vol. 22(3), pp. 313–318 (2003)
16. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., Romeny, B.T.H., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. In: Computer Vision, Graphics, and Image Processing, vol. 39(3), pp. 355–368 (September 3, 1987)
17. Puff, D.T., Pisano, E.D., Muller, K.E., Johnston, R.E., Hemminger, B.M., Burbeck, C.A., McLelland, R., Pizer, S.M.: A Method for Determination of Optimal Image Enhancement for the Detection of Mammographic Abnormalities. Journal of Digital Imaging 7(4), 161–171 (1994)
18. Reeves, T.H., Jernigan, M.E.: Multiscale-based image enhancement. In: IEEE Canadian Conference on Electrical and Computer Engineering, Issue Date: 25-28, vol. 2, pp. 500–503 (1997)
19. Rocchini, D., Di Rita, A.: Relief effects on aerial photos geometric correction. Applied Geography 25, 159–168 (2005)
20. Tzelepis, N., Nakos, B.: A Study on the Lighting Factors affecting Relief Presentation. In: Proceedings of the 21st International Cartographic Conference (ICC), pp. 1343–1350 (2003)
21. Vrhel, M.J., Trussell, H.J.: Filter considerations in color correction. IEEE Trans. Image Processing 3, 147–161 (1994)
22. Yoon, J.H., Ro, Y.M.: Enhancement of the Contrast in Mammographic Images, using the Homomorphic Filter Method. Inf. & Syst. Letter E85–D(1), 298–303 (2002)
23. Young, I.T.: Shading Correction: Compensation for Illumination and Sensor Inhomogeneities. Current Protocols in Cytometry, 1–12 (2000)
24. Yu, W.: Practical anti-vignetting methods for digital cameras. IEEE Trans. on Cons. Elect. 50, 975–983 (2004)
25. Guo, Y., Zhang, X., Zhan, H., Song, J.: A novel illumination normalization method for face recognition. In: Li, S.Z., Sun, Z., Tan, T., Pankanti, S., Chollet, G., Zhang, D. (eds.) IWBRS 2005. LNCS, vol. 3781, pp. 23–30. Springer, Heidelberg (2005)
26. Zheng, Y., Lin, S., Kang, S.B.: Single-Image Vignetting Correction. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 461–468 (2006)
27. Zhu, J., Liu, B., Schwartz, S.C.: General illumination correction and its application to face normalization. In: Proceeding of ICASSP, pp. 133–136 (2003)

Multibeam Echosounder Simulator Applying Noise Generator for the Purpose of Sea Bottom Visualisation

Wojciech Maleika, Michał Pałczyński, and Dariusz Frejlichowski

West Pomeranian University of Technology, Szczecin,
Faculty of Computer Science and Information Technology,
Zolnierska 49, 71-210, Szczecin, Poland
{wmaleika,mpalczynski,dfrejlichowski}@wi.zut.edu.pl

Abstract. Hydroacoustic data form almost the only basis for seabed imaging [1]. The accuracy of scanning devices is therefore crucial for the reliability of the images. In the paper, the precision of the Simrad EM3000 multibeam echosounder was examined by means of statistical methods on real data. The maximum and mean errors were estimated, the correlation of error value and angle between the beam and the vertical line could be observed as well. The results of the experiments will help to increase the accuracy of terrain modeling of the sea bottom what will result in more realistic visualizations.The results of those experiments were implemented in the multibeam echosounder simulator.

Keywords: bathymetric data processing, multibeam echosounder, error of measurement, simulation.

1 Introduction

More than 70% of the Earth's surface is covered by open water. Human life and progress are both strongly connected with it. The analysis of the bottom of the sea or river has become crucial when it comes to safe sailing. Modern imagery techniques significantly enhance efficiency and safety. They are often intended for the representation of the hydroacoustic data, collected by means of SONAR technology, for the analysis of the underwater biological and physical characteristics. Among its important applications is the mapping of the shape and type of marine environments, e.g. shape and bottom type. In order to perform this task more precise methods for scanning, monitoring, and measuring are needed. Nowadays, sonar survey techniques are considered as having more advantages as opposed to traditional methods, because of their rapid and high-resolution work ([2]). Multibeam bathymetry allows to present the seabottom in a convenient and reliable way (see Fig. 1). Usually, the bathymetric data are processed by means of algorithms belonging to various disciplines, e.g. digital signal and image analysis and processing. The Digital Terrain Model (DTM) constitutes their final result.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 285–293, 2011.

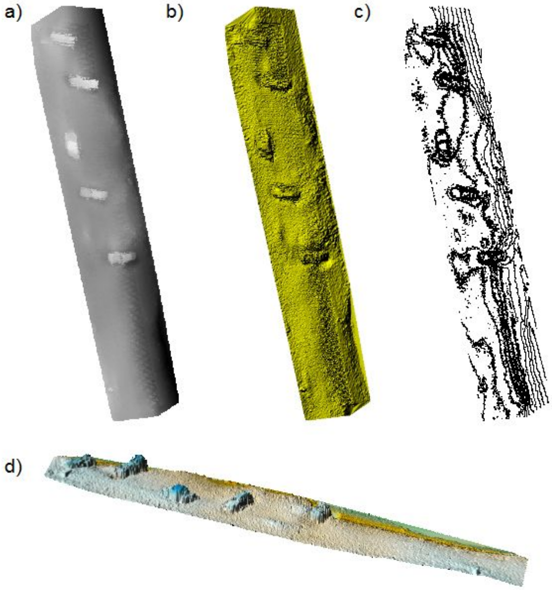


Fig. 1. Examples of DTM (Digital Terrain Model) imagery techniques (cars on the bottom of a river): a) image, b) shaded relief, c) contour map, d) 3D surface

One of the crucial problems encountered in the synthesis of the DTM based on the bathymetric data is the uncertainty of measurement, since an error in measurement may be caused by various factors. This can significantly influence the resultant image representation of the DTM. Although the measurement precision of particular devices is provided by producers, the observed errors are often higher than reference values given by manufacturers.

The problem of measurement uncertainty in case of the bathymetric data is not a new one in scientific literature. In [3] the influence of the including strong uncorrected angular variations in some tracks was stressed. In [2] the variations in the penetration of acoustic beams across different sediment types have been pointed out as the one of the five factors that hamper the proper analysis of the sediment porosity. In [4] the importance of the problem was emphasized, as according to the IHO (International Hydrographic Office) standards a few millimetres accuracy is requested. However, it is often unattainable. In some cases the error can exceed 1 meter ([4]). In [5] and [6] examples of measured uncertainty have also been provided.

The problem of error distribution is very important in the field of echosounder simulation algorithms. The application of noise typical for the measurement device, or even for a specific model, increases the reliability of the generated data.

In the paper, a continuation of research described in [7] is presented, where the authors introduced the algorithm for error estimation in real sonar data.

Here, the distribution of error is investigated. The results of noise application in echosounder simulator are presented as well.

The central problem of this work is the lack of the reference data for error estimation since there are no other sources of information about the seabed.

2 Multibeam Echosounder and Its Simulation

The multibeam echosounder is currently one of the most frequently applied sources of bathymetric data [1]. Its most important parameters are the number of beams and beam angle. The bathymetry, even for small areas results in very large data sets, exceeding millions of points, recorded in the order of acquisition. The principles governing the multibeam echosounder operation cause the points to form lines, which are named here the measurement lines, orthogonal to the vessel's trajectory. An example of such data is shown on Fig. 2.

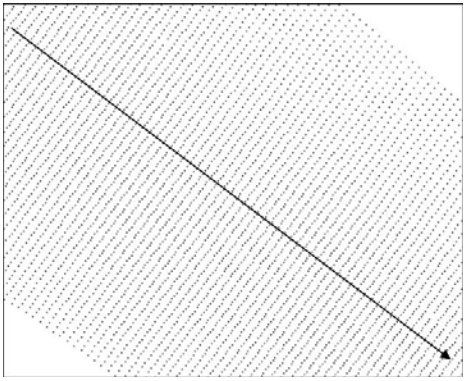


Fig. 2. The orthogonal projection of a set of bathymetric data. Measurement lines typical for multibeam echosounder can be observed. The arrow shows the ship's trajectory.

Distribution and density of bathymetric data makes them completely useless for visualization. Usually, the digital terrain model (DTM) in the form of GRID or TIN is made by means of one of the interpolation algorithms ([1]). The problem of DTM verification has driven the authors of the paper to develop the multibeam echosounder simulator, as the tool for the DTM creation research ([8]). The algorithm of the simulation based on ray-tracing technique was presented in [9]. The simulation would be unreliable if it did not involve the noise generator, but the distribution of measurement errors made by real devices is unknown.

3 Research Procedure

Points, collected in the bathymetric data set, were grouped into consequent measurement lines, which were processed one by one. The basic concept underlying data accuracy investigations in this work was introduced in [7]. The research data set was collected in the area of almost flat bottom and the 10-th degree polynomial approximation of each measurement line was chosen as the reference data for error estimation.

For each point of the measurement line, the difference of depth between the profile and the point was the estimator of the measurement error. Fig. 3 presents examples of measurement lines and corresponding theoretical bottom profiles. Estimated error values were additionally classified depending on the angle between a beam and the vertical line (named here “beam angle”) and error value. Angles between −90 and 90 degrees were grouped into intervals of 10 degree width, error values between 0 and 20 cm were grouped into intervals of 1 cm width. The algorithm of the data analysis is shown on Fig. 4.

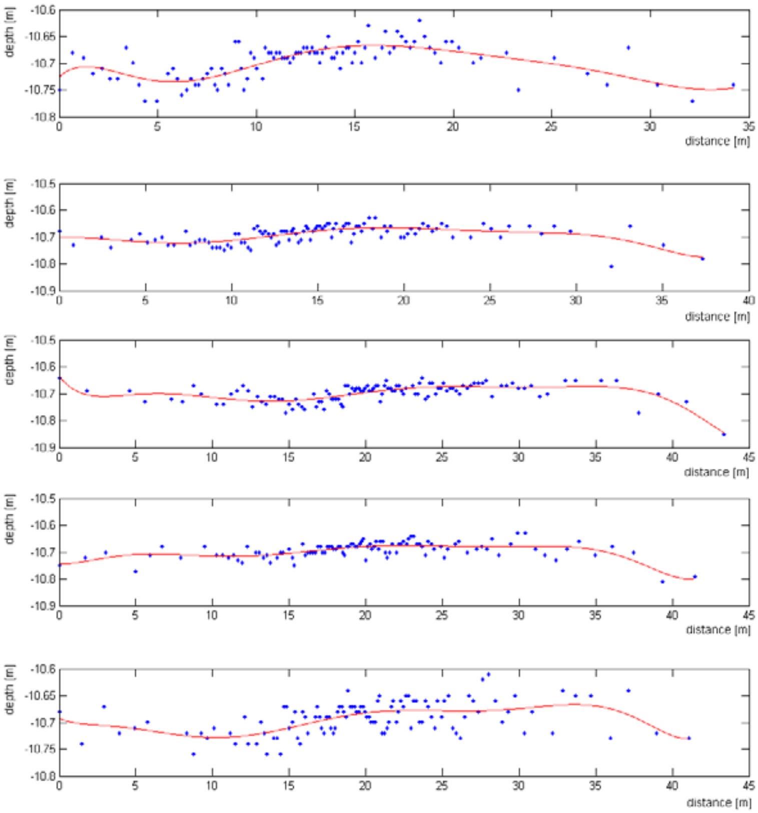


Fig. 3. Examples of measurement lines and corresponding theoretical bottom profiles

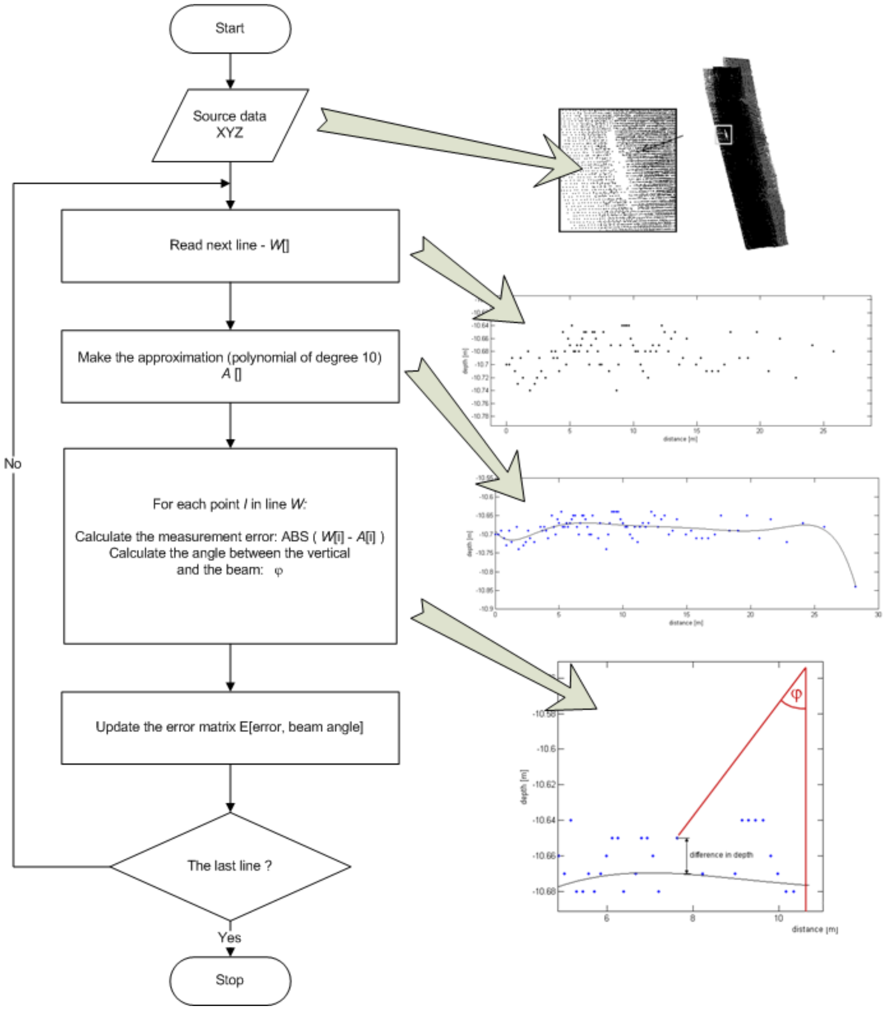


Fig. 4. The applied algorithm of the data analysis

4 Results

The data for the research was collected by means of the multibeam echosounder Simrad EM3000 from part of the Pomeranian Gulf (Baltic Sea). The surveyed area had the shape of a rectangle of the size 1500 × 1400m, and had an almost flat bottom surface the depth of which was approximately 10m. The data set consisted of approx. 17 millions of points in 156 thousands of measurement lines. By means of the algorithm shown in the previous section the measurement errors were estimated, for each interval of angles separately. The quantities of errors of different values were also calculated. The analysis of the results leads to consideration of nearly random distribution of measurement errors in each measurement line. The dependence of the error quantities on error values seems to follow Gaussian distribution. Number of measurements a few centimeters deeper and shallower than the theoretical bottom profile was very similar, which is not dependent on the angle of beam. The distribution of error values depending on the angle of beam is presented in Fig. 5.

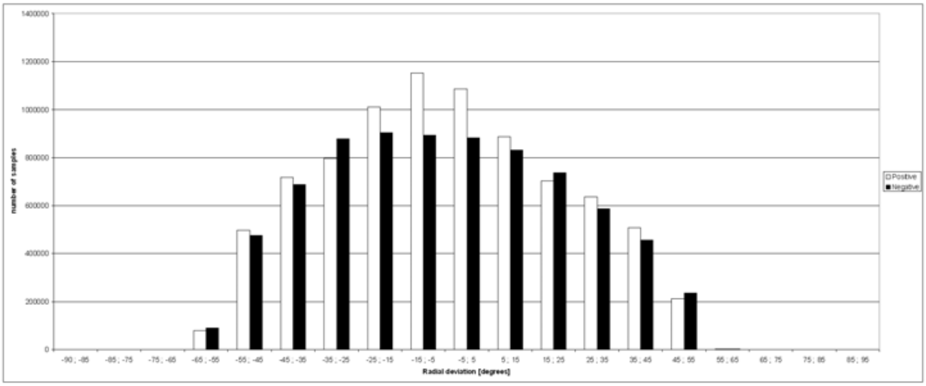


Fig. 5. The number of measurements deeper and shallower than the theoretical bottom profile for each beam angle

The average error value for all measurements was 2.25 cm (0.22% of the average depth). For 95% measurements the errors fell below 5 cm and for 99.7% measurements they did not exceed 7 cm. The mean square error was equal to 3.12 cm. The magnitudes of errors slightly increased with the beam angle. For measurements of beam angles 0 – 5 degrees, the mean error was 1.9 cm, while in case of the biggest angles (from 55 to 65 degrees), the mean error was 3 cm (over 50% increase). The distribution of mean error magnitudes depending on the beam angle is presented in Fig. 6.

The standard deviation was also calculated for each beam angle interval (Fig. 7). Its value increases with the beam angle (from 0.1 cm for the vertical measurements to 0.42 cm for the biggest angles). The results fulfill the natural assumption about decrease of measurement accuracy witch increase of beam angle.

5 Application of Obtained Results in the Echosounder Simulator

The results of experiments helped to improve the multibeam echosounder simulator by implementing a noise generator simulating real distribution of measurement errors. Next, the virtual survey was done over DTM prepared using real data. Examples of virtual measurement lines are shown on Fig. 8.

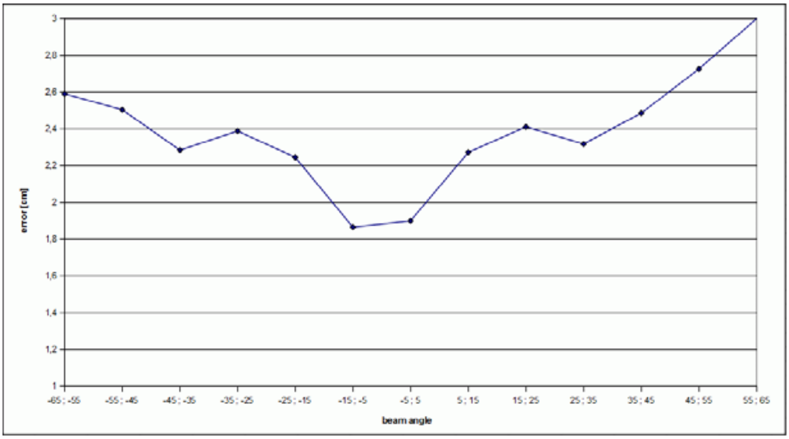


Fig. 6. Magnitudes of average errors for each beam angle interval

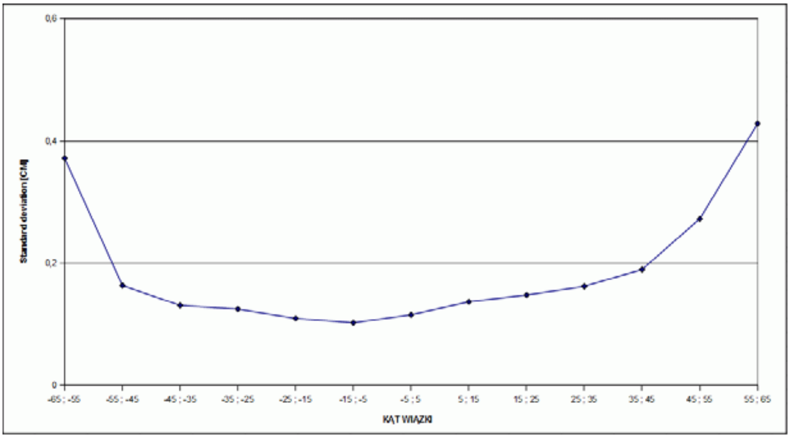


Fig. 7. Standard deviations for each beam angle interval

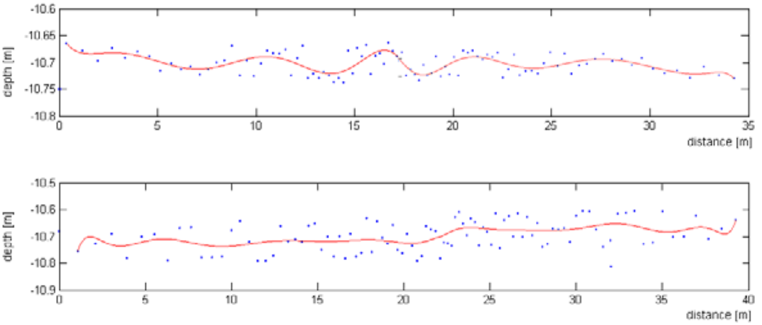


Fig. 8. Examples of virtual measurement lines synthesized using the noise generator based on the research results

Another virtual survey was done without and with noise generating and two DTMs were created basing on both data sets. The difference between the bottom images of those models can be observed on Fig. 9.

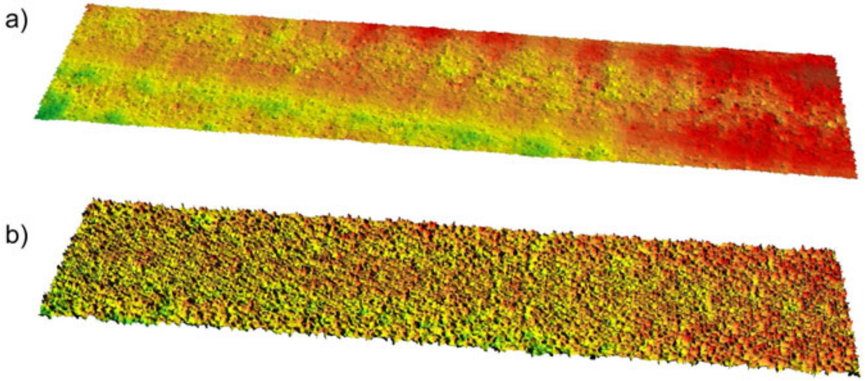


Fig. 9. 3D surfaces of digital terrain models created basing on virtual survey data a) noise generator off, b) noise generator on

6 Conclusions

The analysis of the data collected using the Simrad EM3000 multibeam echosounder has shown that the measurement errors are random and increase with the beam angle. The average error value 3 cm does not significantly differ from the error declared by the producer (RMS=5cm [10]). However, it should be noted that the measurements were performed on small depths. The analysis of data has shown a 50% increase of error magnitudes for measurements of the same line, depending on the beam angle.

The distribution of measurement errors estimated in the research was implemented in the noise generator of the multibeam echosounder simulator resulting in the improvement of the accuracy of DTM verification algorithms.

In the future, similar analysis of various measurement devices made by different producers could be performed. The impact of depth on the measurement error should be examined.

Acknowledgment. This work has been supported by the Ministry of Science and Higher Education of the Republic of Poland under grant no. N N526 073038.

References

1. Stateczny, A. (ed.): The methods of the comparative navigation. Scientific Association of Gdansk (2004) (in Polish)
2. Sutherland, T.F., Galloway, J., Loschiavo, R., Levings, C.D., Hare, R.: Calibration techniques and sampling resolution requirements for groundtruthing multibeam acoustic backscatter (EM3000) and QTC VIEW(TM) classification technology. Estuarine, Coastal and Shelf Science 75(4), 447–458 (2007)

3. Blondel, P., Gomez Sichi, O.: Textural analyses of multibeam sonar imagery from Stanton Banks, Northern Ireland continental shelf. Applied Acoustics 70(10), 1288–1297 (2009)
4. Le Bas, T.P., Huvenne, V.A.I.: Acquisition and processing of backscatter data for habitat mapping — Comparison of multibeam and sidescan systems. Applied Acoustics 70(10), 1248–1257 (2009)
5. Brown, C.J., Blondel, P.D.: Developments in the application of multibeam sonar backscatter for seafloor habitat mapping. Applied Acoustics 70(10), 1242–1247 (2009)
6. Ferrini, V.L., Flood, R.D.: A comparison of Rippled Scour Depressions identified with multibeam sonar: Evidence of sediment transport in inner shelf environments. Continental Shelf Research 25(16), 1979–1995 (2005)
7. Maleika, W., Pałczyński, M., Frejlichowski, D., Stateczny, A.: Analysis of survey data collected using Simrad EM3000 multibeam echosounder. Methods of Applied Computer Science (4/2010), 55–64 (2010) (in Polish)
8. Maleika, W., Pałczyński, M.: Virtual multibeam echosounder in investigations on sea bottom modeling. Methods of Applied Computer Science (4/2008), 111–120 (2008)
9. Maleika, W., Pałczyński, M.: Development of Virtual Multibeam Echosounder. Accepted for Publication in WAT Bulletin, vol. 1 (2011) (in Polish)
10. http://www.kongsberg-simrad.de/pdf/faecherlot_em3000_eng_broch.pdf

Automatic Segmentation of Digital Orthopantomograms for Forensic Human Identification

Dariusz Frejlichowski and Robert Wanat

West Pomeranian University of Technology, Szczecin,
Faculty of Computer Science and Information Technology,
Zolnierska 52, 71-210, Szczecin, Poland
{dfrejlichowski,rwanat}@wi.zut.edu.pl

Abstract. Dental radiographic images are one of the most popular biometrics used in the process of forensic human identification. This led to the creation of the Automatic Dental Identification System with the goal of decreasing the time it takes to perform a single search in a large database of dental records. A fully automated system identifying people based on dental X-ray images requires a prior segmentation of the radiogram into sections containing a single tooth. In this paper, a novel method for such segmentation is presented, developed for the dental radiographic images depicting the full dentition — pantomograms. The described method utilizes the locations of areas between necks of teeth in order to determine the separating lines and does not depend on the articulation of gaps between adjacent teeth, thus improving the results achieved in the situation of severe occlusions.

Keywords: image segmentation, dental pantomography, dental human identification, ADIS, forensic identification.

1 Introduction

Human forensic identification is the process of determining the identity of a person using the evidence available in legal investigations. Teeth and bite are particularly popular for this application ([1]), as they are both robust to decomposition and highly discriminant ([2]). In the wake of ever growing use of digital radiology, the Federal Bureau of Investigation (FBI) created the Dental Task Force (DTF) to improve the use of digital dental information in legal proceedings ([3]). One of its main tasks was the creation of an Automated Dental Identification System (ADIS) in order to preliminarily browse through a large database of dental X-ray images in the search for radiograms with the most similar characteristics to present in an input image. The features used for comparison in ADIS are morphologic, e.g. the shapes of individual teeth or their dental restorations. The model and functionality of ADIS were described in [4].

In the simplified model of ADIS ([5]), there are three preliminary steps before two radiograms are compared: image enhancement, image segmentation (separation of the image into regions containing at most one tooth) and feature

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 294–302, 2011.

extraction (finding the contour of the tooth). In this paper, a novel method for segmentation of panoramic dental X-rays is proposed and presented. Panoramic X-rays, or pantomograms, are a type of extraoral radiograms (i.e. radiograms, where the film is located outside of the patient's mouth) showing the full dentition on a single image. A sample pantomogram (all pantomograms in this paper are presented courtesy of Pomeranian University of Medicine in Szczecin) can be seen in Fig. 1. To our best knowledge, few segmentation algorithms were developed solely for the use with pantomograms (see Section 2). Panoramic images contain the largest amount of information about dentition, but due to the rendering 3-dimensional, semi-circular geometry of the jaw onto a 2-dimensional image, the teeth presented on pantomograms have a tendency of occluding with each other.

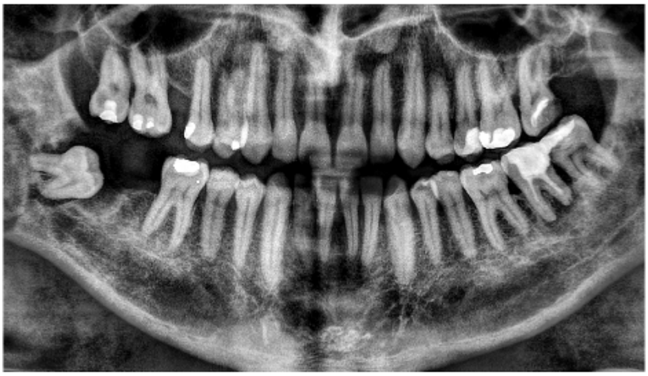


Fig. 1. A sample pantomogram after image enhancement and cropping

2 Existing Segmentation Methods

As has been mentioned in the previous section, there are several existing methods for dental radiogram segmentation. Most of these methods have been created with intraoral images in mind.

The first method was presented by Jain and Chen in [6]. It is focused on the application of the integral projections of pixels for the detection of gaps between teeth. The algorithm is separated into two parts: the first is the detection of the gap between lower and upper jaw, and the second is the detection of gaps separating individual teeth. The former step needs user input, so it can be considered semi-automatic. After the initial point of the gap has been selected by the user, moving in both directions, the algorithm chooses short horizontal lines with the highest probability of belonging to the gap. The probability is calculated using the equation ([6]):

$$p_{v_i}(D_i, y_i) = p_{v_i}(D_i) p_{v_i}(y_i), \quad (1)$$

where $p_{v_i}(D_i)$ is the normalized integral projection of a given horizontal line subtracted from 1 and $p_{v_i}(y_i)$ is a Gaussian with expected value equal to the position of the last chosen line (or the user selection in the first iteration). This probability function has its maximum for the horizontal line that is vertically close to the last selected line and that is composed of pixels with low values. After the maxima have been found for the whole image, a spline function is applied to form a smooth line that becomes the separating line between upper and lower jaw. Once the spline has been calculated, for every point on the curve a new integral projection is calculated in the direction perpendicular to its local curvature. These projections obtain low values in areas between teeth, thus the search for gaps can be reduced to searching for valleys in the plot of the integrals. The areas between these three lines (gap between upper/lower jaw, two successive gaps between teeth) and the horizontal borders of the image become the segments used later in the process of feature extraction. An improved version of this method was proposed in [7]. In this case, the images are processed through wavelet kernels before calculating integral projections to further accentuate the gaps between the necks of teeth.

Another method, presented in [8], consists of the use of active contour models, also known as snakes. Initially described in [9], snakes are a model of parametrized splines driven towards edges and lines on the image by external forces, i.e. forces derived from the image on which they operate, as well as internal forces, i.e. user imposed control over the elasticity and rigidity of the contour. In [8], the external driving force along the contour of the snake E_{ext} is given as:

$$E_{ext}(x, y) = G_\sigma(x, y) * I(x, y), \tag{2}$$

where G_σ is a 2-dimensional Gaussian and $I(x, y)$ is the original image. Thus defined external force amounts to the intensities Gaussian-filtered original image, in which case it takes the lowest values in the dark areas of the radiogram, such as the gap between upper and lower jaw or in spaces between teeth. When using properly selected initial approximations of these curves it achieves very good segmentation of the image.

3 Description of the Proposed Method

3.1 Preliminary Steps

For the proper work of the proposed method it is assumed that before an image is segmented using the method, it had been enhanced using the algorithm presented in [10]. That method is based on the decomposition of the image into a Laplacian pyramid, separating the radiogram into smaller images containing progressively lower frequencies of the signal present in the original image. Then, a range of simple filters is applied to selected layers of the pyramid, including sharpening filter and contrast enhancement methods, before the image is recomposed again. It is also necessary to locate the gap between the frontal teeth before the segmentation. In this paper, a nose position detection is used and then a vertical line on the same position is considered the center.

3.2 Separating the Upper and Lower Jaw

The first step is the determination of the line separating the upper and lower jaw. The same method as presented in [6] and briefly described in the previous section is used. In order to automatize the process of segmentation, instead of requiring that the user inputs the initial separating point used by the algorithm, it is selected by choosing the horizontal integral projection around the center of the image with the lowest value, usually between 40% and 60% of the height of the image. Since the teeth on the image create an arc, instead of using the full horizontal line that would pass through teeth further from the incissors, only a small number of pixels (equal to 20% of the width of the image) closest to the previously selected frontal teeth gap is used to calculate the projections. The mentioned paramaters were established experimentally. Afterwards the algorithm proceeds as described in [6].

3.3 Localization of the Areas between the Necks of Teeth

The obtained curve is then used to estimate the position of the neck of every tooth that is the part of the tooth where roots end and the formation of crown and enamel begins. While the crowns of separate teeth tend to occlude with each other and roots are difficult to separate from the underlying bone, the area between the necks of two adjacent teeth is distinguishable enough to be easily found on a pantomogram. Since the necks of teeth are on the same height as dental pulp, which is darker than the surrounding teeth, the simplest method for finding a line going through dental necks is to translate the line separating upper and lower jaw vertically, sum the intensities of pixels the line passes through for every translation and select the ones for which there is a distinctive drop of values, indicating the line passes through darker areas of the pulp.

In order to guarantee that neither the original gap between jaws nor a gap between roots of teeth and the edge of the image are selected in lieu of the desired dental pulp line, the vertical translation scope should be chosen to conduct the search for a limited range, automatically discarding translations too close and too far from the line separating jaws. The result of this step are two values, one negative and one positive, that indicate the amount of pixels that the vertical position of every point belonging to the spline separating jaws should be moved for in order to receive the spline that passes through dental pulps of teeth in each jaw.

The next step is the selection of points on each spline representing a gap between the necks of two adjacent teeth. To refine the results of this stage of the algorithm, a new image is created by multiplying the value-inverted original image with local range filtered version of the original image. That image has high values for darker pixels that lie in areas with neighboring points of low and high intensity. For both upper and lower jaw, an array of values is saved containing the intensities of points belonging to the splines passing through their respective dental pulps. Sharp spikes on the plot of these values indicate dark spots surrounded by light regions, indicating a gap between necks of teeth. In order to remove false spikes, the values of the function are smoothed using the Gaussian filter. Then, starting from the previously selected line separating frontal teeth, small subsets of the values in the array are chosen for comparison. To determine the size of these subsets, average widths of teeth on every position were calculated based on twenty sample pantomograms. Both jaws are fairly symmetrical considering the size of teeth on a given position, thus only 8 values need to be calculated for each jaw, one for every tooth from first incissor to last molar.

To select the proper spike value indicating the gap between the necks of teeth for a currently searched tooth on the position p_c, a Bayesian probability $P(x_i, p_c)$ is calculated for each point on the curve (x_i) according to the equation:

$$P(x_i, p_c) = I(x_i)G(x_i, p_c)D(x_i, p_c), \tag{3}$$

where I is the intensity of the range filtered and inverted original image in the point x_i and G is a discrete Gaussian function with the expected value equal to the horizontal position of the last detected gap displaced left or right (depending on the current search direction) by the amount of pixels equal to the average width of a tooth on position p_c. Thus, if the last detected gap lied on position 100 and the average width of the currently searched tooth is 50 then the expected value should lie on position 50, if the search direction is left, or 150, if the search direction is right. D is a function introduced to narrow the amount of pixels considered during every iteration of the algorithm, equal to 1 for points in the horizontal distances from the previous detected gap between 75% and 175% of the expected width of a tooth on position p_c, with the regard to the current search direction, and equal to 0 elsewhere. During every iteration the current argument maximum of the equation 3 is added to the list of gap positions and becomes the starting point of the next iteration of the algorithm. If the amount

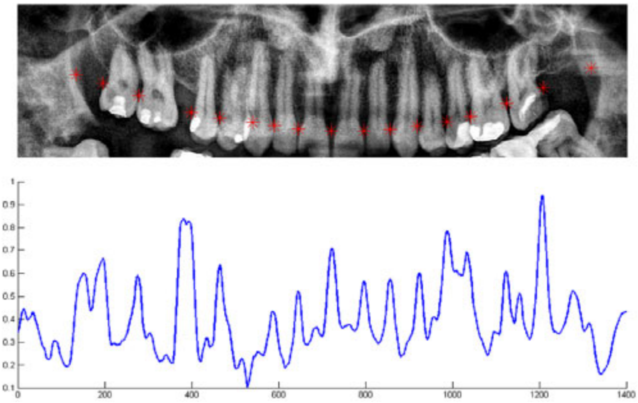


Fig. 2. Located gaps between necks of teeth for the upper jaw of the sample pantomogram (top) and the corresponding values of the pixels on the range filtered and inverted original image through which the dental pulp spline passes (bottom)

of gaps in a given search direction for either upper or lower jaw equals 8 or the vertical edge of the image was reached, the algorithm stops. The result of this step of algorithm for the sample pantomogram is presented in Fig. 2.

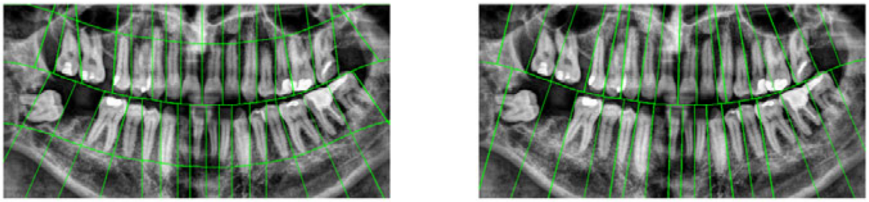
Thus calculated gap locations provide a good estimation of the position of areas between teeth. However, in some cases a simple vertical line is not sufficient for the separation of two adjacent teeth. Molars and, in some cases, premolars require an additional step of the algorithm to determine the angle of the segmenting line. In order to find a straight line seperating these teeth, an additional point needs to be found between them. A greedy algorithm was used, iteratively moving 1 pixel towards the top or bottom of the radiogram, choosing the pixel in horizontal vicinity with the highest intensity on the inverted and range-filtered image and using it as the basis for the next iteration. After the number of iterations equal to half of the length of an average tooth on a pantomogram, the position of the last result becomes the second separating point. The line passing through the first and second separating point becomes the segmentation line.

3.4 Removing the Areas Below the Roots of Teeth

The last step is to remove the areas below the roots of teeth. It is similar to the detection of dental pulp line, i.e. the curve separating both jaws is translated vertically in search of an alignment where the sum of pixels it passes through is lower than the surrounding results, indicating that the area between the teeth line and the cheekbone line was achieved. The only difference between this part of the algorithm and the search for the line of necks of teeth is the range of translations considered during the search. After finding separating lines, every segment of the image lying between the jaws gap line, two consecutive lines separating adjacent teeth and the line below the dental roots is considered an area possibly containing a tooth and is later used in the process of feature extraction.

4 Results and Discussion

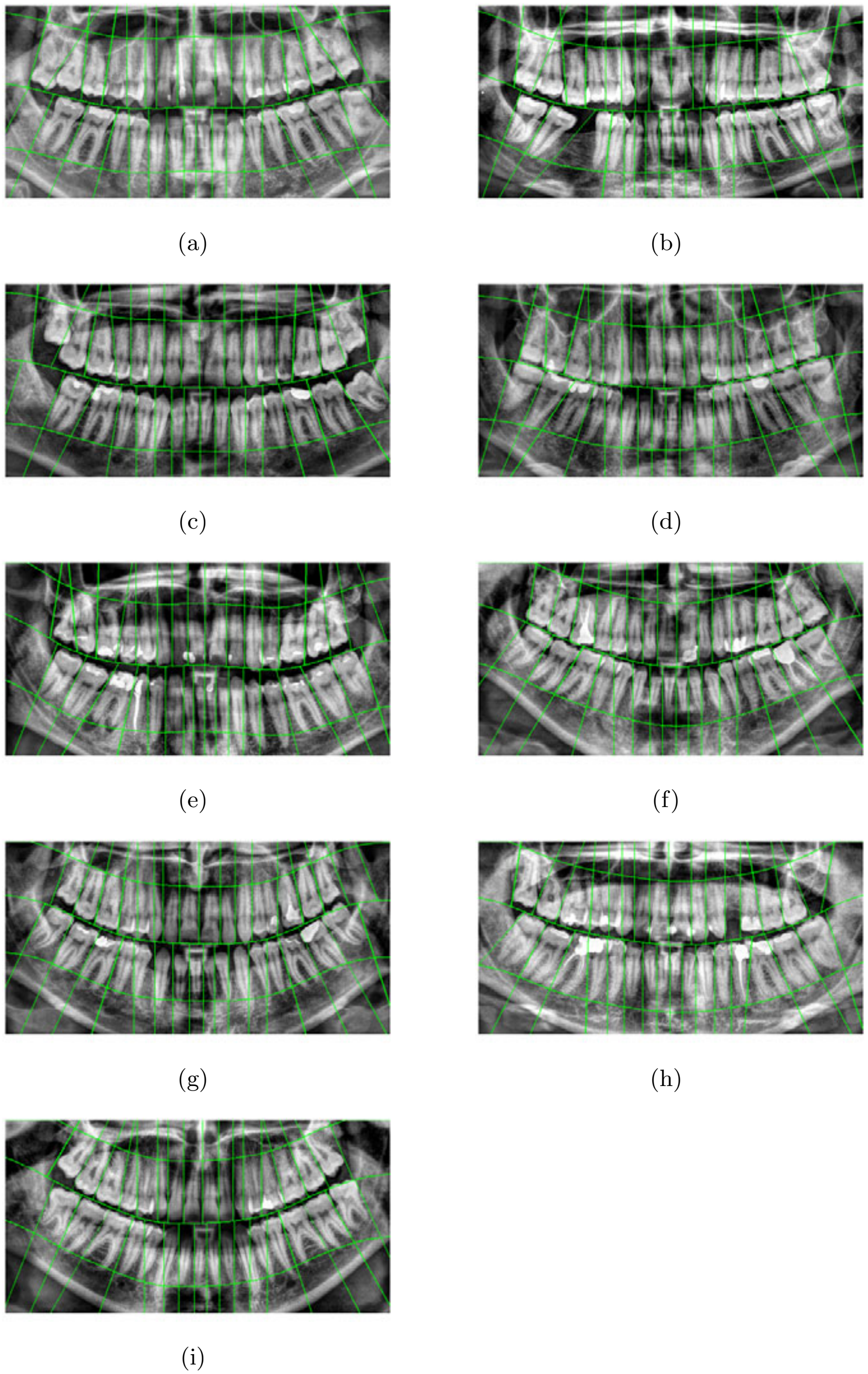
The described method was tested on a database containing 218 orthopantomograms and some exemplary test results are presented in this section. Firstly,



(a)

(b)

Fig. 3. A comparison of the results abtained for the method proposed in the paper (3(a)) and the integral projections method described in [6] (3(b))



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Fig. 4. Exemplary test results obtained using the presented method

Fig. 3 presents a comparison between the results of the proposed method and the most popular approach applied so far — the integral projections ([6]). The results are comparable, with a slightly better result achieved for the proposed method — smaller number of lines passing through dental roots and the removal of the area below the roots of teeth make the segments more compact and improve the possibility of a correct shape recognition in the next step of the system. It should also be noted that during the tests the calculation of the function seen in the bottom of Fig. 2, used to detect spikes in the values indicating the gaps between teeth, took less than 1 second, whereas the calculation of the analogous function for the integral projections, i.e. the $p_{v_i}(D_i)$ in equation 1 took around 212 seconds. After these calculations, both methods use the same algorithm to detect the spikes in the resulting plot, as described in equation 3.

Further results were presented in Fig. 4. It can be seen that the segmentation algorithm provides satisfactory results. If the teeth can be separated using a single line, the algorithm is usually able to find the optimal line of separation. The described method is also able, in some cases, to separate unerupted teeth, as can be seen in Fig. 4(c). The obvious bad results can be found in cases of occlusions so severe that it is impossible to find a straight line to separate both teeth, such as in the case of the incissor and the first premolar (respectively third and fourth tooth from the center of the image) of Fig. 4(a) and Fig. 4(e). Malaligned teeth can also be separated to a degree, but in the case of severe problems in alignment the separating line can not capture the whole tooth within a segment, for example the furthermost bottom left molar in Fig. 3(a). Detection of the ends of dental roots helps in removing bright areas of the underlying bone that would be otherwise attributed to the tooth, but in some cases a fragment of the tooth is removed as well, e.g. in Fig. 4(e). The last problem stems from the fact that dental pulp can easily be mistaken for the edge of the tooth, resulting in a mis-segmentation. This happened in the case of the lower left canine in Fig. 4(b), where the separating line passes through the middle of the tooth, but because average teeth widths are used, the final 8th tooth is separated correctly.

5 Summary and Conclusions

In this paper a novel method for segmenting dental panoramic radiograms into regions containing single teeth was described. It uses a different approach than the existing algorithms developed for intraoral images, focusing on detecting gaps between necks of teeth and roots of teeth, that are both easy to find on a pantomogram and allow to separate teeth even in the presence of occlusions.

The method provides satisfactory results, however some changes can be introduced in the future. A considerable improvement of its results could be achieved if a more sophisticated method was used instead of the greedy algorithm to determine the second point through which the separating line between two teeth is traced. The method could also be used as an initial step for further segmentation using for example active contours, what can improve the obtained results.

References

1. Bowers, M.C.: Forensic Dental Evidence. Elsevier, Amsterdam (2004)
2. Lee, S., et al.: The Diversity of Dental Patterns in Orthopantomography and its Significance in Human Identification. Journal of Forensic Science 49(4), 784–786 (2004)
3. Nassar, D., Ammar, H.H.: A Prototype Automated Dental Identification System (ADIS). In: Proceedings of the 2003 Annual National Conference on Digital Government Research, pp. 1–4 (2003)
4. Abdel-Mottaleb, M., et al.: Challenges of Developing an Automated Dental Identification System. In: IEEE Mid-West Symposium for Circuits and Systems, Cairo, Egypt, pp. 411–414 (2003)
5. Fahmy, G., Nassar, D.E., Haj-Said, E., Chen, H., Nomir, O., Zhou, J., Howell, R., Ammar, H.H., Abdel-Mottaleb, M., Jain, A.K.: Towards an Automated Dental Identification System (ADIS). In: Zhang, D., Jain, A.K. (eds.) ICBA 2004. LNCS, vol. 3072, pp. 789–796. Springer, Heidelberg (2004)
6. Jain, A.K., Chen, H.: Matching of Dental X-ray Images for Human Identification. Pattern Recognition 37(7), 1519–1532 (2004)
7. Said, A., et al.: Dental X-ray Image Segmentation. In: SPIE Technologies for Homeland Security and Law Enforcement Conference (2001)
8. Zhou, J., Abdel-Mottaleb, M.: A Content-based System for Human Identification Based on Bitewing Dental X-ray Images. Pattern Recognition 38(11), 2132–2142 (2005)
9. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active Contour Models. International Journal of Computer Vision 1(4), 321–331 (1988)
10. Frejlichowski, D., Wanat, R.: Application of the Laplacian Pyramid Decomposition to the Enhancement of Digital Dental Radiographic Images for the Automatic Person Identification. In: Campilho, A., Kamel, M. (eds.) ICIAR 2010, Part II. LNCS, vol. 6112, pp. 151–160. Springer, Heidelberg (2010)

Common Scab Detection on Potatoes Using an Infrared Hyperspectral Imaging System

Angel Dacal-Nieto[1], Arno Formella[1], Pilar Carrión[1],
Esteban Vazquez-Fernandez[2], and Manuel Fernández-Delgado[3]

[1] Computer Science Department, Universidade de Vigo,
Campus As Lagoas 32004 Ourense, Spain
angeldacal@uvigo.es

[2] GRADIANT, Galician R&D Center in Advanced Telecommunications, Spain

[3] Centro de Investigación en Tecnoloxías da Información (CITIUS),
Universidade de Santiago de Compostela, Spain

Abstract. The *common scab* is a skin disease of the potato tubers that decreases the quality of the product and influences significantly the price. We present an objective and non-destructive method to detect the common scab on potato tubers using an experimental hyperspectral imaging system. A supervised pattern recognition experiment has been performed in order to select the best subset of bands and classification algorithm for the problem. Support Vector Machines (SVM) and Random Forest classifiers have been used. We map the amount of common scab in a potato tuber by classifying each pixel in its hyperspectral cube. The result is the percentage of the surface affected by common scab. Our system achieves a 97.1% of accuracy with the SVM classifier.

Keywords: Hyperspectral, Infrared, Potato, SVM, Random Forest.

1 Introduction

Detecting and identifying defects and diseases in potato tubers (*Solanum tuberosum*) continue to be an important challenge for food engineering and automation. Industry uses a large variety of technologies and computer vision methods have been a specially successful choice. Nevertheless, some new technologies should be taken into account for improving non-destructive potato quality assessment.

The importance of the potato industry is extreme, since potatoes are still one of the most consumed products in the world; they are the world's fourth largest food crop. The annual production is 325 million tons and it moves an amount of global transactions of about 6 billion US dollars (2007 data). Thus, the world potato average consumption is 31 kg per capita and year [1].

Hyperspectral imaging is an emerging technology originally designed for military remote satellite inspection [2], but also used for remote sensing, astronomy and earth observation. It is also a reliable approach to classical spectroscopy, because despite a little loss of accuracy, an object can be analysed in significantly less time, in a non-destructive way.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 303–312, 2011.

The scientific community has started to show its interest in the last years in hyperspectral imaging possibilities for food quality [3]. Regarding the research in potato quality assessment, there are systems to predict the water content in potatoes using classical spectroscopy techniques [4]. Some other contributions are oriented to the detection of clods between a set of potato tubers using hyperspectral imaging [5]. Finally, there are contributions [6] that investigate composition characteristics from potato tubers like water, starch and proteins, using invasive spectroscopy techniques, meanwhile others [7] use NIR spectroscopy to predict specific gravity and dry matter in potatoes. Using other optical spectral methods [8], there are contributions for the detection of common scab, dry rot, gangrene, and other diseases, using wavelength ranges between 590 nm and 2030 nm, and getting accuracies up to 83%. Unfortunately, these systems are either destructive or they can not be easily included in classical machine vision developments in order to use the same image acquisition for all the processes.

Our objective is to map the common scab affected areas in potato tubers. This has been achieved in the past by using different technologies, as it has been described before. However, mapping the common scab is not only a required objective itself: we also use this mapping as a preprocessing stage into a wider potato inspection system, which detects internal and external defects and diseases. Some of these diseases require a morphological study, so hyperspectral technology seems to be the best approach. It would be interesting to provide a solution using the same image acquisition system, in order to unify the inspection process, so hyperspectral imaging has been also the selected technology to solve the common scab mapping problem.

Our solution is objective, automatic and non–destructive. Nevertheless, this choice makes difficult the comparison with other common scab detection methods. In fact, there are not hyperspectral solutions yet for detecting common scab, due to the novelty of the technology. Moreover, previous spectral contributions used different wavelength ranges, or performed a combined searching of other diseases, so a partial comparison is presented.

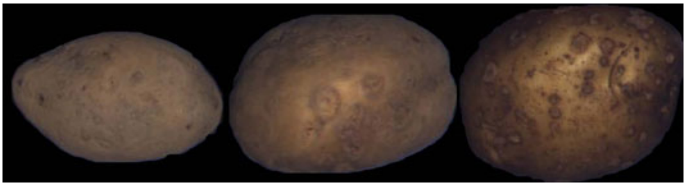


Fig. 1. Three examples of common scab affected potatoes

2 Image Acquisition System

The concept of hyperspectral imaging is to perform a spectroscopic analysis of the light reflected or transmitted by the object of interest. We couple a spectrograph and a matrix camera to obtain both spectral and spatial information.

The camera is a Xenics Xeva 1.7–320 with an InGaAs 320 × 256 pixel sensor and USB connection ("http://xenics.com”); the spectrograph is a Specim Imspector N17E ("http://specim.fi”). Both are sensitive from 900 nm to 1700 nm. The system has also three 50 W AC halogen lamps placed in the inspection plate. The illumination is diffused by the reflection in a plastic dome over the plate.

The spectrograph has a linear input (one pixel height), where the x-axis represents the same x-axis (spatial) of the object. The y-axis (spectral) is then *studied* to obtain how every pixel in the row varies along the spectral range.

With one spectral image, we are inspecting only one spatial line, so we need to perform the inspection over the whole object. This is accomplished by joining a rotatory Mirror Scanner ("http://specim.fi”) to the spectrograph. It is based on performing the mirror rotation over the object, taking care of synchronization between mirror stepping and image acquisition (Figure 2). Finally, the obtained images are transposed in order to obtain the hyperspectral cube (Figure 3).

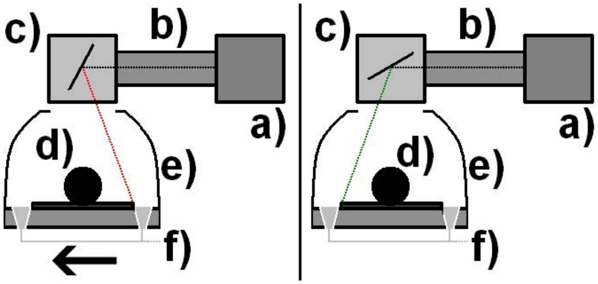


Fig. 2. Left: scanning initial position at 70°. Right: scanning final position at 110°. The arrow shows the direction of scanning. Hyperspectral system scheme: a) camera, b) spectrograph, c) mirror scanner, d) object, e) diffuse chamber, f) halogen lamp.

To sum up, our system obtains 320 spectral images (320 × 240 pixels), that are transposed into hyperspectral cubes formed by 256 images with 320 × 320 pixels, corresponding to 256 consecutive wavelengths from 900 nm to 1700 nm.

3 Experiment

We use a set of 234 potato tubers (variety Agria) from Xinzo de Limia (Spain), with different degrees of common scab incidence, that have been collected from some potato packing companies during the 2009 harvest.

3.1 Segmentation

In every hyperspectral cube, we need to segment the potatoes from the background for later mapping tasks. We segment only one image from the hyperspectral cube (the wavelength 980 nm has been found after several performance tests). We obtain a mask that is applied in the rest of the cube.

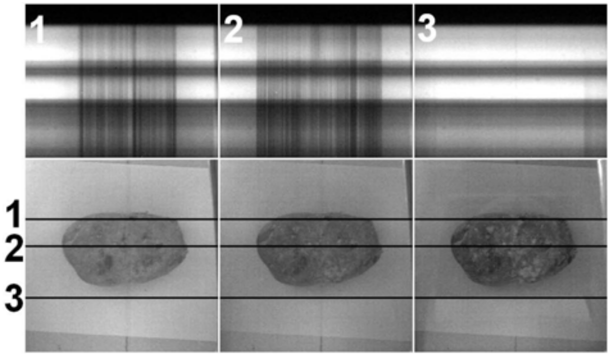


Fig. 3. Up: Three spectral images taken from different lines of the object. Down: 978 nm, 1173 nm, and 1608 nm spatial images.

Segmentation runs in several steps (Figure 4), helped by the open source library OpenCV [9]. First, we binarize the image using Otsu's method [10] that calculates the optimum binarization threshold using a probabilistic analysis of the image. Then, a Gaussian blurring clusters the noise in the image. Another binarization is needed before a connected-component labelling, performed to remark contiguous areas in the image. At this point, we know that the blob with the largest area (excluding the background) is the potato. We select this blob and create the mask used to segment all the images in the hyperspectral cube.

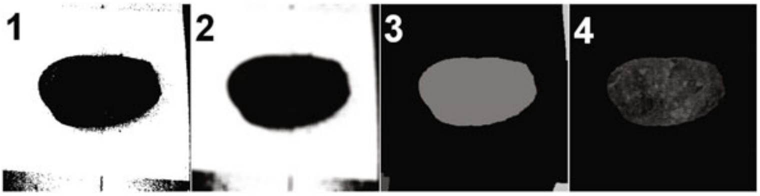


Fig. 4. 1: Binarization using Otsu's method. 2: Smooth operation. 3: Blob analysis. 4: Example image after applying the full mask.

3.2 Feature Extraction

In our problem, we have to distinguish two classes: *common scab* and *healthy*. To create a dataset with both common scab affected samples and healthy samples, experts helped us to identify which portions (ROI) were affected and which were not. Some hyperspectral cubes provided more samples (especially those more affected by common scab) meanwhile others provided just one healthy sample.

Note that every hyperspectral cube consists of 256 images that correspond to the 256 bands of the hyperspectral system. For this reason, when we select a ROI, we are not selecting just a rectangle of pixels, but that rectangle all over the 256 images that are part of the hyperspectral cube.

When selecting the ROI, the average intensity value of the pixels in the ROI is calculated for each band. Hence, every sample (independently of its size) is represented with 256 attributes and an extra attribute that denotes the class (common scab or healthy). Eventually, we have obtained 649 samples (208 corresponding to common scab class and 441 corresponding to healthy class).

The samples can be visualized in a chart where the x-axis represents the wavelength range and the y-axis the grey level in the band, which is actually the arithmetic mean of pixels in every band of the ROI. In the Figure 5 we can see ROI selection of two samples and the corresponding luminance charts.

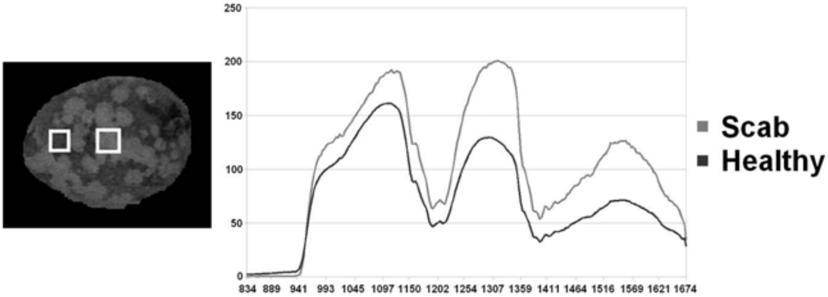


Fig. 5. Left: healthy and common scab affected (the brightest) ROI's. Right: Luminance charts from two different samples. The x-axis represents the wavelength. The y-axis represents the average grey level in the ROI, for each band.

3.3 Feature Selection

Feature selection is a common task in pattern recognition, specially if the initial number of features is high. With less features, the learning process is faster and the generalization capabilities of the classifier are improved. In our case feature selection is a fundamental step to decrease the overall execution time, identifying which wavelengths are sufficient to solve the common scab detection problem.

We have tested some techniques regarding spectral bands selection on hyperspectral imaging systems, implemented on Weka [11]: Genetic Search [12] (which selects 11 bands), Scattered Search [13] (11 bands), Greedy Stepwise [14] (5 bands), Linear Forward Selection (LFS) [15] (7 bands), and Correlation-based Feature Subset Selection (CFS) [16], (6 bands). Note that with CFS, three contiguous zones have been selected: 1300 nm–1303 nm, 1336 nm–1342 nm and 1503 nm. Techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are the most commonly used unmixing techniques in spectral imaging. However, these algorithms do not reduce the number of wavelengths needed, rather they generate a linear combination of the 256 features into a new feature space. This is the reason why only feature selection operations are interesting in this research.

To summarize, after this step of the experiment, we provide six datasets to the classification procedure: *genetic*, *scattered*, *greedy*, *LFS*, *CFS* and *full*.

3.4 Classification Algorithms

We present results for two classification algorithms: Random Forest (RF) and Support Vector Machines (SVM). Other classifiers have been tested in a preliminary stage (Logistic Regression, MLP and k-NN), but they performed poorly.

A Random Forest [17] is a collection of trees that classifies individually an input sample, and then evaluates the individual responses of the trees to output the mostly voted class. We used the OpenCV implementation of RF, tuning the m_{try} parameter, which is the number of features to be used in random selection.

SVM find the optimal hyperplane over a high dimensional space where the feature vectors have been mapped using a kernel function (Gaussian in our case). We can tune its behaviour with the regularization parameter (also known as cost, or C), which is not very relevant for the results [18], and the kernel spread (γ), with high relevance on the classification accuracy. SVM has been introduced in our system with the library LibSVM [19].

3.5 Classification Evaluation Procedure

For each dataset, we evaluated the classification algorithms using a method based on randomly generating 10 permutations of the dataset, so that each permutation has the same samples, but differently ordered. Then, each permutation is divided into three parts: *training* (50% of the samples), *validation* and parameter tuning (25% of the samples), and *test* (remaining 25%). The samples are normalized (zero mean and standard deviation one) to avoid that attributes in greater numeric ranges influence excessively over those with smaller variation.

For each combination of tunable parameters and for each permutation, we train a classifier using the training sets. Then, we test its performance by using the validation sets. We selected the parameter values which provide the best average accuracy over the 10 permutations.

In the case of RF, the default value of m_{try} is $\sqrt{p}$, being p the number of features of the problem. We follow a parameter tuning as being suggested by [20]. We use different values of m_{try}: $m_{\text{try}} = p^0$, $m_{\text{try}} = \sqrt{p}$, $m_{\text{try}} = p/4$ and $m_{\text{try}} = p/2$. The rest of the parameters have been established as [20] recommends. Thus, the number of trees has been set to 500, since it is enough, and there is no penalty for having an excessive number of trees.

In the case of SVM, we try pairs of (C, γ) using exponentially growing sequences for C and γ [19]. Thus, we use $C = 2^n, n = -5..14$ and $\gamma = 2^n, n = -15..3$, which gives 380 combinations. A finer adjustment has been discarded after some preliminary tests.

Finally, for each permutation, we train the classifier using the training sets tuned with the best parameters found, evaluating its accuracy on the test sets.

Note that using more permutations prevents unfair divisions of the dataset. For example, using only one permutation, if all the *easy-to-classify* samples are filled in the test set, it would cause unfairly good results. Additionally, each dataset has also been evaluated using leave-one-out cross-validation (loocv).

3.6 Affected Surface Measurement

Once being able to classify, the objectives are, for each potato (actually for each hyperspectral cube), obtaining an image that marks which zones are affected and computing the percentage of common scab affected surface.

First, we segment the potato, removing the background from the hyperspectral cube. Then, each pixel in the hyperspectral cube is classified individually (excluding the background, that has been localized previously in the segmentation step). Each hyperspectral pixel has (in our system) 256 values: however, depending on the feature selection procedure, we will have only a few of them. With the information of membership of each pixel to one class or another, we create a common scab map image. To reduce noise in the final map, a closing operation is performed followed by an opening operation with the same kernel. Finally, we calculate the percentage of the affected surface.

The objective of our system is to inspect 20 Kg samplings. Each potato will be inspected only by one side. Inspecting such an amount of potatoes averages individual errors, since we provide a statistical measurement. The result will be the average affected surface in the whole 20 Kg sampling.

4 Results and Discussion

The results of all datasets and classifiers can be seen in Table 1 including the leave-one-out cross-validation. Support Vector Machines show to be more effective than Random Forest in all the datasets. On the other hand, the CFS dataset seems to have the better subset of features, so that the pair SVM and CFS dataset is the best option to solve our problem.

Table 1. Accuracy (in %) for each dataset and classification algorithm

Classifier	Dataset	Accuracy %	loocv Acc. %	Best params.		Valid. Acc. %	Bands
					m_{try}		
RF	full	95.4	96.1		$\sqrt{p}$	95.6	256
	genetic	94.3	96.2		$p/2$	93.6	11
	scattered	93.8	96.9		$p/4$	95.0	11
	greedy	95.9	97.4		$p/2$	96.7	5
	LFS	94.5	96.8		$\sqrt{p}$	95.2	7
	CFS	95.8	96.5		$\sqrt{p}$	96.1	6
				C	γ		
SVM	full	96.2	96.6	2^{-2}	2^{-10}	96.4	256
	genetic	96.0	96.8	2^{5}	2^{-10}	96.5	11
	scattered	95.7	96.9	2^{5}	2^{-2}	96.5	11
	greedy	96.7	96.9	2^{12}	2^{-15}	97.0	5
	LFS	96.0	97.7	2^{7}	2^{-1}	96.9	7
	CFS	**97.1**	**98.0**	2^{11}	2^{-5}	**97.4**	6

As commented in Section 3.3, PCA and similar methods have been analysed but considered not adequate. However, some preliminary work has been done to check their performance. Thus, a new dataset has been created using Weka, after applying the PCA method to the *full* dataset. The loocv results show that this dataset gets a 95.2% of accuracy with RF, and approximately a 96% of accuracy using SVM. These results are 2 points under the feature selection algorithms results, and even worse the *full* dataset.

Now we are going to study further the best dataset–classifier pair. The best combination of parameters found was to be $C = 2^{11}$ and $\gamma = 2^{-5}$. The confusion matrix can be seen in Table 2. Note that these results were obtained using the test sets, composed by 25% of the samples (162 in our case). This is an average confusion matrix taking into account the ten permutations.

Table 2. Average confusion matrix obtained with the CFS dataset using SVM

Classified as *Real*	Common Scab	Healthy
Common Scab	48.3	3.1
Healthy	1.6	109

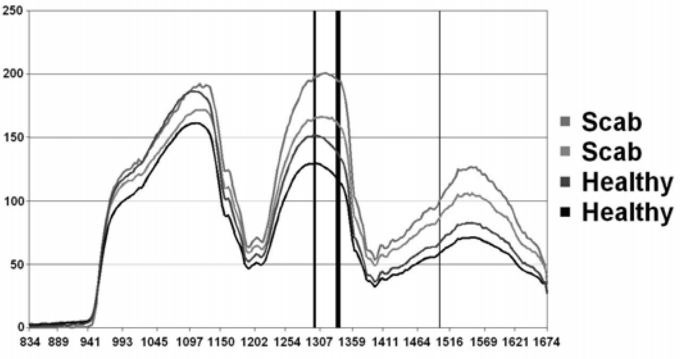


Fig. 6. Four samples, two from each class. Columns in grey mark the zones used by the CFS dataset. The rest of the bands were not selected. The x-axis represents the wavelength. The y-axis represents the grey level.

Four samples (two of each class) are presented in their luminance chart in Figure 6. Columns in black mark zones being selected in the CFS dataset. Previous contributions [21] show that at wavelengths greater that 1100 nm, where absorption by water dominates, the reflectance increases due to dehydration in the affected area, as in the case of common scab. Our automatically selected wavelengths lie in that range. However, by the moment it is impossible to compare our results with other common scab detection methods, since they use different wavelength ranges, or searched for other diseases in the same experiment.

5 Conclusions

Hyperspectral imaging has shown to be an good technology applied to food quality assessment. We have used an objective and non-destructive infrared hyperspectral system to identify the surface affected by common scab on potatoes.

Several feature selection algorithms have been tested, showing that this is a critical step to increase the system speed, because only 6 bands achieve the best accuracy. The selected bands with the CFS method (1300 nm, 1303 nm, 1336 nm, 1339 nm, 1342 nm and 1503 nm) provide enough information to classify common scab and healthy surface with a 97.1% of accuracy using the SVM classifier (tuned with $C = 2^{11}$ and $\gamma = 2^{-5}$).

This information could be useful for the designing of a specific multispectral image acquisition system, which would not have any mechanical device to move the camera or the object, because images could be captured within a reduced and specifically chosen group of wavelengths (in our case around 1301 nm, 1339 nm and 1503 nm). Hyperspectral cube reconstruction would not be needed any more. Hence, the time spent in an acquisition session would be considerably reduced.

The system will be used as a preprocessing step to remove the common scab for improving other disease identification algorithms on potatoes, as hollow heart, or the dry matter estimation amount. In future work, it would be interesting to evaluate the system with other potato varieties. On the other hand, methods like LDA should be tested in order to compare with the feature selection methods used in this paper. Finally, the relationship between the wavelengths selected with the best dataset and the biological components of common scab should be researched. This could be achieved by using a different image acquisition system (i.e. sensitive from 500 nm to 2000 nm), in order to compare our results with the obtained in [8].

Acknowledgements. This work was partially supported by the Spanish regional government of the "Xunta de Galicia" (projects PGIDIT08TIC004CT and PGIDIT08MMA010402PR). We want to acknowledge the "Laboratorio Oficial de Metroloxía de Galicia (LOMG)" and the "Centro Tecnolóxico da Carne (CTC)" for their helpful collaboration.

References

1. Potato World - International Year of the Potato (2008), http://www.potato2008.org/en/world/index.html (accessed January 01, 2011)
2. Goetz, A., Vane, G., Solomon, J.E., Rock, B.N.: Imaging spectrometry for earth remote sensing. Sci. 228(4704), 1147–1153 (1985)
3. Sun, D.: Hyperspectral Imaging for Food Quality Analysis and Control. Academic Press, Elsevier, San Diego, California (2009)
4. Singh, B.: Visible and near-infrared spectroscopic analysis of potatoes. M.Sc. Thesis. McGill University, Montreal, PQ, Canada (2005)
5. Al-Mallahi, A., Kataoka, T., Okamoto, H., Shibata, Y.: Detection of potato tubers using an ultraviolet imaging-based machine vision system. Biosyst. Eng. 105, 257–265 (2009)

6. Buning-Pfaue, H.: Analysis of water in food by near-infrared spectroscopy. Food Chem. 82, 107–115 (2003)
7. Kang, S., Lee, K., Son, J.: On-line internal quality evaluation system for the processing potatoes. In: Food Process. Autom. Conf. Proc., Providence, Rhode Island (2008)
8. Porteous, R.L., Muir, A.Y., Wastie, R.L.: The Identification of Diseases and Defects in Potato Tubers from Measurements of Optical Spectral Reflectance. J. Agric. Eng. Res. 26, 151–160 (1981)
9. Bradski, G., Kaehler, A.: Learning OpenCV: Computer Vision with the OpenCV Library. O'Reilly Media, Sebastopol (2008)
10. Otsu, N.: A threshold selection method for gray level histograms. IEEE Trans. Syst. Man Cybern. 9, 62–66 (1979)
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1) (2009)
12. Goldberg, D.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading (1989)
13. García-López, F., García-Torres, M., Melián-Batista, B., Moreno-Pérez, J.A., Moreno-Vega, J.M.: Solving feature subset selection problem by a Parallel Scatter Search. Eur. J. Oper. Res. 169(2), 477–489 (2008)
14. Weihs, C.: Multivariate exploratory data analysis and graphics, a tutorial. J. Chemom. 7, 305–340 (1993)
15. Guetlein, M., Frank, E., Hall, M., Karwath, A.: Large Scale Attribute Selection Using Wrappers. In: Proc. IEEE Symposium on Computational Intelligence and Data Mining, pp. 332–339 (2009)
16. Hall, M.: Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand (1998)
17. Breiman, L.: Using Iterated Bagging to Debias Regressions. Mach. Learn. 45, 261–277 (2001)
18. Valentini, G., Dietterich, T.G.: Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods. J. Mach. Learn. Res. 5, 725–775 (2004)
19. Chang, C.C., Lin, C.J.: LIBSVM:a library for support vector machines (2008), `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`
20. Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P.: Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. J. Chem. Inf. Comput. Sci. 43, 1947–1958 (2003)
21. Gunasekaran, S., Paulsen, M.R., Shove, G.C.: Optical methods for nondestructive quality evaluation of agricultural and biological materials. J. Agr. Eng. Res. 32, 209–241 (1985)

Automatic Template Labeling in Extensible Multiagent Biometric Systems

Maria De Marsico[2], Michele Nappi[1], Daniel Riccio[1], and Genny Tortora[1]

[1] Biometric and Image Processing Laboratory, University of Salerno,
Via Ponte Don Melillo, 84084 Fisciano (Salerno) Italy
[2] Department of Computer Science, Sapienza University of Rome,
Via Salaria 113, 00198 Rome Italy
demarsico@di.uniroma1.it, {mnappi,driccio,tortora}@unisa.it

Abstract. Many works in literature have demonstrated the superiority of multibiometric systems compared to single-biometrics ones, in terms of both accuracy and robustness. Most current multibiometric systems implement a static architecture, which does not change in time. However, the ability to progressively add more modules, either to process more biometrics or to exploit additional algorithms, might contribute to further enhance recognition performance. The addition of a new module (agent) to an already fully operational multiagent system usually requires its preliminary setup and training. In particular, it must be provided with a brand-new gallery, whose templates are suitably labeled according to the represented identities; alternatively, an existing database of templates, formerly built according to the suited feature extraction procedure, might be updated to include better quality items. It would be of paramount importance if the new agent can "inherit" the "experience that was already acquired by the other agents, including the creation of its gallery without having to undergo a full enrolling phase in its turn. We present here an algorithm to align a new module to the already existing ones in an automatic and unsupervised way. Experimental results show that our algorithm is effective both when the new database must be created from scratch (sample labeling), as well as when it is pre-existing and must be updated (sample updating). The latter operation can also be iteratively performed in running modules to dynamically update their galleries. In particular, we present here results achieved for face recognition.

Keywords: Multibiometric system, template labeling, unsupervised learning, template updating.

1 Introduction

Two issues related to biometric systems, in particular to multibiometric ones, are attracting research attention: a) template labeling, where the gallery of a new added module is created from scratch, and labels must be assigned to each template to assign it to a certain identity [7], and b) template updating, where old as well as corrupted templates for an identity are substituted by more recent or representative ones [8]. The

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 313–322, 2011.

presence of more subsystems can represent an important added value in defining effective solutions for both these problems. Differently from other template updating approaches (see for example [6]), our proposal substantially relies on the widely recognized ability of multibiometric systems to provide better recognition accuracy, thanks to the joint contribution of either different biometrics, or modalities, or algorithms. In general, the presence of more modules either implies an external coordination, to merge their results and update their state, or requires that the modules themselves have such built-in ability. The first solution allows to better exploit off-the-shelf software solutions for the single biometrics. Many architectural choices are available in literature, to combine different biometric systems. For this work, we chose N-Cross Testing architecture with Supervisor Module (NCT-SM) [1], thanks to its extensibility. The N-Cross Testing (NCT) architecture, considers a multibiometric system as a multiagent one. Each agent is an autonomous and replaceable unit, with a possibly different weight on the final response. For instance, in the version of NCT implemented herein, for each recognition operation, the identity returned by each agent is associated with a reliability measure (we use System Response Reliability – SRR [3]), which expresses how much we can trust each single response. Information from all agents is suitably fused to formulate the global response (Figure 1).

When a new agent is to be integrated into an existing NCT architecture, its gallery is empty and must be populated from scratch through a suitable enrolling phase; moreover, agents galleries have to be updated from time to time to maintain the more representative templates. The approach we propose allows to avoid a new ad-hoc enrolling phase, and to exploit the responses of the other NCT systems instead, to "guide" the new agent in building its database until it can be considered as fully operational. We can consider our automatic labeling as a form of auto-training and unsupervised learning (since no operator intervention is required). We will also sketch how the approach can be extended to implement template updating in a multiagent system. Of course, our approach is not appropriate for a typical watch-list identification, were the aim is to exclude that the probe coincides with one of relatively few individuals (e.g. terrorists, or missing people). The probability of a positive identification in this case is too low to rely on it in order to populate the involved agents' galleries. On the other hand, it is worth mentioning the typical category of situations were our approach is feasible. We consider a kind of closed set identification (re-identification) which requires to continuously check that the subjects within a protected area (a strategic chemical farm, or a flight deck) are all and only among those authorized to stay there. The difference with the classical closed set identification is the addition of an alarm if the returned identity is not enough similar to the probe at hand. Of course, it would be impractical to periodically perform such re-identification through explicit identity verification (with identity claim by each subject at hand). Several classifiers concur to determine the identity to label/update without human intervention, but rather borrowing from the compound system. In general, template updating is an aspect of biometric systems on which research is recently focusing, but "the state of the art related to template update is still in its infancy" [6]. NCT architecture and the labeling (template updating) algorithm have been tested according to some of the protocols suggested in [6]. We randomly selected the subjects to submit to the system and evaluated the performance variations against the number of updates.

2 The N-Cross Testing Protocol

2.1 Agent Reliability

Biometric agents might not be equally reliable, e.g. due to different performances of classifiers, or to intrinsic characteristics of biometrics at hand. Moreover, not all responses from the same agent are equally trustworthy, e.g. due to unstable input devices. Therefore, we adopt the System Response Reliability (SRR) index [3], a system/gallery dependent metric that can assess the ability of a biometric system to identify an enrolled subject, for each single probe. Two versions of it are introduced in [3], and they both exploit the full list of gallery subjects, ordered by similarity with the probe, which is returned by the identification module. In a few words, the version of SRR that we use for this work is computed considering how much "crowded" is the cloud of different individuals (each individual may correspond to more templates) which are returned in the similarity ordered list after the first one. The radius of the cloud is determined experimentally, and the more crowded the cloud, the less reliable the response (the higher the possibility of a wrong answer). Each agent T_k returns, for each of its responses $s_{k,i}$, $i=1,\ldots$, a *reliability measure* $srr_{k,i}$ in the range [0, 1] (the higher, the better). SRR can even be computed when using an off-the-shelf biometric application, given that this returns an ordered list of (all) candidates in the gallery in response to each identification operation. See [3] for further details. Each T_k is characterized by an estimated threshold th_k, such that a response $s_{k,i}$ is reliable only if $srr_{k,i} \geq th_k$. This threshold is different from the acceptance one, which would normally regulate acceptance or rejection based on a similarity/distance measure. In fact, we may obtain an acceptance response which is poorly reliable, as well as a rejection which is fully reliable. Therefore, the overall acceptance rule which results from using SRR requires that reliability of any response is controlled first. If this is sufficient, acceptance or rejection are decided according to a similarity/distance threshold. Otherwise, a repetition of the identification operation may be requested. Depending on the security requirements of the system, the reliability threshold th_k can start from an initial value of 0 (all responses are considered as reliable) and be updated over time to reach its more suitable value. Otherwise, a better initial value can be computed during a system tuning phase. In both cases, in the proposed system, the threshold of each agent T_k can be further updated over time, thanks to a supervisor module that coordinates the single biometric applications [1].

2.2 System Architecture

NCT architecture fits either multibiometric systems (each agent processes a different biometrics), or multimodal ones (all agents process the same biometrics, captured under different modalities), or multiexpert ones (the agents process the same biometrics, under the same modality, yet extracting different features). In the present case, we consider a multiexpert system on the face biometrics [1]. Each agent processes the same face image from which different kinds of templates are computed. Each template feeds the appropriate agent. In basic NCT, N agents T_k, k=1, 2,…, N, mostly work in parallel, but exchange information at some points to reach the final result (cross operation) (Figure 1). Each agent has a database (gallery) G_k of biometric templates $t_{k,j}$, j=1, 2,…, |G_k| which are labelled with the enrolled identities. The exchanged items are the respective lists of templates (galleries) ordered by similarity with the probe. In [1] different versions of the basic procedure are presented. For readers' convenience, we report here some details of the one adopted for this work, which exploits the SRR index discussed above. When the j-th probe p is submitted to the global system, each agent T_k independently extracts its template p_k, compares it with those in its specific gallery, and produces its ordered list of subjects. Each agent also computes a reliability measure $srr_{k,j}$ for such response. Such value is used to determine how the agent participates to the second phase of the identification process (cross). As discussed in the previous section, each agent is characterized by a reliability threshold th_k: it sends its candidate list to the others only if its own SRR index for the current response is above such threshold. The sent list is truncated to the first M subjects, with M fixed in advance and experimentally determined. This step starts the cross phase, where each T_k first merges the lists from (reliable) companions (not including its own one, except when it is the only reliable system, as discussed below). Returned scores are suitably normalized in order to consistently merge information from different agents. Unreliable agents still work, even if they do not send their list, by receiving and merging lists from reliable companions. Each agent produces a merged list, with identities only coming from reliable ones. Unreliable agents will all produce the same list, i.e. the one obtained by merging all the reliable ones. Thus, in the special case when only one agent is reliable, all the unreliable companions return its list, and, since it does not receive any list form the others, it returns its own list as well. The length of each merged list depends on the amount of overlap, since shared subjects get a single average score. Finally, each T_k returns the first subject in its (resorted) merged list, its score and a variation of SRR index. The new reliability measure assigned to the first identity in each merged list L_k, is given by the average of reliability indexes of (reliable) agents in the set LS_k contributing to the list:

$$LSRR_k = \frac{1}{|LS_k|} \sum_{T_i \in LS_k} SRR_i \tag{1}$$

In this way the result returned by each T_k is influenced by the other agents. The decision phase produces the final response. In the present proposal, a *Supervisor Module (SM)* [1] (Figure 1) exploits single agent responses and their reliability to compute the final global response. It also evaluates the overall system state and updates its parameters. This takes to a self-tuning and co-evolution of agents.

The identity receiving the majority of votes is possibly returned as the input subject (see below). If more identities get the same number of votes, the basis for the final choice is not the score but the LSRR index. The probe is recognized if the final score got by the candidate identity (the mean of those from voting agents) complies with a threshold δ. Our SM updates single reliability thresholds according to the behaviour of the companion agents. Agents agreeing on the correct answer are rewarded by a lower threshold (their responses will be more easily accepted by the system), while disagreeing agents are charged by increasing their thresholds (in future operations, they will need to score a higher reliability for their answers to be accepted).

The process converges to an optimal threshold configuration independently from the starting one. More details on experiments supporting such claim can be found in [1]. In the present proposal, as a variation of the behaviour presented in [1], SM module also drives the setting up of newly added agents, as shown below.

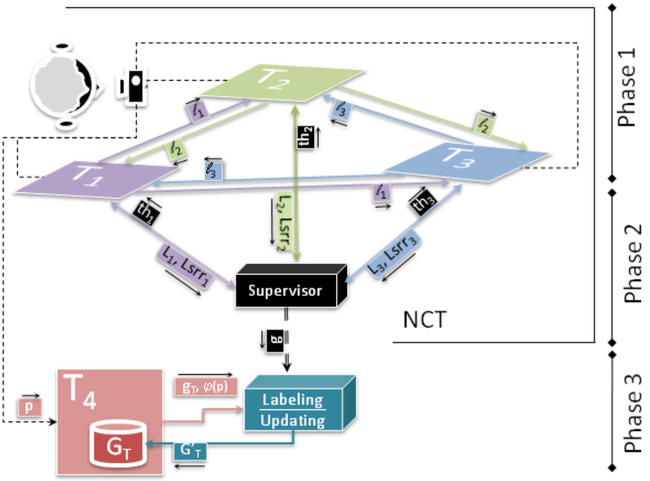


Fig. 1. Functional schema of the proposed system. In the first phase (first level) the existing agents T_1, T_2 e T_3, exchange their respective lists l_k and compute the merged list with global LSRR. In the second phase (second level) such information is sent to the Supervisor Module which computes the global response (identity g), and updates and transmits the new SRR thresholds to all three agents T_1, T_2 e T_3. It also transmits the global response to the Labeling/Updating algorithm (third level), which uses it to update the gallery of T_4.

3 Labeling Algorithm

Due to its highly modular nature, a NCT system can be possibly expanded with new agents, given they are trained to handle the current working setting. This implies that they must undergo a training phase to set up their feature extraction process (consider for example PCA or LDA). Afterwards, their gallery must be populated according to a specific biometric trait of registered users. In practice, the gallery must be filled with templates, and these must be labeled according to enrolled individuals. In our case, also the computation of the reliability of the new agent must be initialized. In this section we discuss how to perform these two latter activities through an automatic process. It is worth noticing that the exploited algorithm can be applied either when a gallery is empty, and the new inserted templates must be labelled, or when, during normal operation, it contains elements to be updated. Notice that template substitution and/or deletion may occur either during the initial labelling phase, or during updating, as will be discussed later. If deletion during update is allowed, and the gallery only contains one template per identity, a situation may occur with missing information about one or more subjects. While this might bias the final result, we have methods in literature to address this problem [4]. The main strength of the algorithm is the use of

SRR returned values, which drive gallery labelling/updating. Given *PR* a set of probes, *G* a gallery of templates, *I* the set of registered identities, *SRR_values* the reliability values in the interval [0, 1], *SIM_values* the similarity values in the interval [0,1], and *SIM_threshold* the acceptance threshold based on similarity, we first define some functions: $id{:}G \rightarrow I \cup NULL$ is such that $id(t)$ is the identity corresponding to the template *t* (*NULL* if *t* is not labelled yet); $temp{:}I \rightarrow P(G)$ is such that $temp(x)$ is the subset of templates in *G* which correspond to identity *x*; *link_id*$:G{\times}I \rightarrow P(G)$ is such that *link_id(t, x)* inserts t to the set of templates corresponding to *x*, i.e. assigns the identity label *x* to *t*; *identify*: $PR \rightarrow G \times$*SRR_values* is such that *identify(p)* is its closest template in *G* with an associated reliability measure; *sim*: $PR \times G \rightarrow \times$*SIM_values* is such that *sim(p,t)* is the similarity value between *p* and *t*. For simplicity, we assume from here on that a gallery contains one template per identity at any moment. We first give the pseudo-code of the labelling algorithm.

```
Labelling pseudocode for T

Input: the current gallery G and related identities I, the
current probe p and a (reliable) retrieved identity (label) g
from the global NCT system

if (g∉I) then
   G = G∪p; I = I∪g; link_id(p,g);
else
   (ct,srr) = identify(p)
   if (id(ct)≠g) OR (id(ct)=g AND sim(p,ct)<SIM_threshold)
      G = G\ct; I = I\id(ct); /*discard template and
                                       associated identity*/
   else             /*id(ct)=g AND sim(p,ct)>SIM_threshold)*/
      G=G\{ct}∪{p}; (ct1, srr1) = identify(ct);
      if (id(ct1)≠g) OR (id(ct1)=g AND
                                    sim(ct,ct1)<SIM_threshold)
         G=G\{p}∪{ct};
      else if (id(ct1)=g) AND (ssr1>srr) NOP; /*gallery
                                                is changed*/
      endif
endif
```

The process works as follows. The new (not yet operative) agent T_k receives the current (reliable) retrieved identity (label) *g* from the global NCT system, and checks if a pertaining template is already present in its own gallery G_k. If not, the current probe template p_k is inserted, and associated (labeled) with the new identity. Otherwise, if *g* is already associated to some template in the gallery, T_k validates it. To this aim, it performs a recognition operation on p_k to obtain the closest template in G_k and a value $srr_{k,p}$. The obtained template ct_k will pertain to a certain identity g_k. We have two possible situations. If the identity g_k is different from the *g* previously returned by the fully operational global system, or even if its similarity with the probe is below the acceptance threshold, both template ct_k and identity g_k will be deleted. In this way, we eliminate a source of possible recognition errors. Consider that the deleted identity and a pertaining template might be possibly re-added in a future step. If this does not happen, and if a specific problem arises for the missing identities (we experimentally found that they would be very few), it will be possible to force the insertion of good quality templates for them, in a significantly reduced mini-enrollment step. This is a little price to pay compared with the decreased performances due to "bad" templates. Alternatively, if g and g_k are the same, T_k starts a second (cross) recognition step. It temporarily changes the roles of probe and gallery templates: the new recognition operation is performed using ct_k as probe, and the gallery $G'_k=G_k-\{ct_k\}\cup\{p_k\}$. A new identity h_k will be retrieved, and a value $srr_{k,ct}$. will be associated to the new response. If the two recognition operations retrieved different identities, i.e. $h_k \neq g_k$, or even if the second recognition falls below the acceptance threshold, the original gallery G_k is restored by re-inserting ct_k and deleting p_k. This is because the original template fostered a correct identification, while the new one caused a recognition error, since it better matches a different (wrong) identity. Otherwise, if $h_k=g_k$, and $srr_{k,ct}>srr_{k,p}$, G'_k becomes the new gallery for T_k, i.e. the just acquired template p_k replaces the former one for the identity g_k, since it fosters a better discrimination from the other identities in the gallery (a better value of *SRR* is obtained when p_k is a gallery template than when it is used as a probe). The reliability threshold for the new agent will start from 0 (all responses acceptable) and will be updated at any recognition operation by the SM, in the same way used for the fully operating agents (see previous section). The agent can start full cooperation with the others when its gallery becomes substantially stable during a sufficient time elapse.

The extension to $l>1$ gallery templates per identity is straightforward. Moreover, a template updating process can be implemented by each fully operative agent if it periodically executes a similar algorithm, possibly except for the deletion of identities in case of discordant results.

4 Experimental Results

The architecture that we set up to evaluate the proposed algorithm included four classifiers, each implementing a different technique for face biometry, and a supervisor module. Three out of the four agents (T_1, T_2, T_3) were already working (fully labelled gallery and determined SRR thresholds) and respectively exploited Linear Discriminant Analysis (LDA), Orthogonal Locality Preserving Projections (OLPP) and Neighborhood Preserving Embedding (NPE) (see 1). The fourth agent T_4, which was added later and which gallery had to be labelled, was based on Partitioned Iterated Function Systems (PIFS) [2]. The sets of face images making up the testing benchmark were extracted from AR Faces database [5], since a good number of subjects (70 men and 56 women) with a sufficiently varied set of distortions is represented in such database. Each subject appears in two different sessions with 13 image sets each. Sets differ in expression (1 neutral, 2 smile, 3 anger, 4 scream), illumination (5 left light, 6 right light, 7 all side light), presence/absence of occlusions (8 sun glasses, 11 scarf), or combinations (9 sun glasses and left light, 10 sun glasses and right light, 12 scarf and left light, 13 scarf and right light). Sets 14 to 26 of the second session present the same conditions at a different time. Performances were measured in terms of Recognition Rate (RR), Equal Error Rate (EER), Percentage of Templates into the Gallery (PTG) defined as $|G_T|/|G|$, where G_T is the gallery of a single system and G is the ideal gallery including the full set of enrolled identities (PTG=1 for fully operational agents). To show how the performances of agent T_4 improved as its gallery was gradually populated/updated, it was evaluated separately from the others at regular intervals. Every 20 operations (insert, delete, substitute), the values for RR, EER, and PTG were computed for T_4 using set 6 as probe, with its current gallery. Set 1 was the gallery for T_1, T_2, T_3 in all experiments; the probe set depended on the experiment at hand, and probes were randomly extracted from it. An experiment ended when performances of T_4 in that setting became stable.

In the first experiment, the initial gallery of T_4 was empty, and the probe set for all four agents was the union of sets 2 and 3. Figure 2(a) always shows PTG<1, i.e. $|G_T|<|G|$, with $|G_T|$ referring to T_4, because some responses from the NCT system (T_1, T_2, and T_3) were marked as unreliable, and therefore the corresponding identities were not inputted to the algorithm for T_4. If we relaxed this policy, more identities might have entered the gallery in less time, but we would have had a more error-prone behaviour from the new agent. Since the gallery for T_4 was initially empty, we can notice a faster convergence towards the final EER and RR. This is because many identities were created directly in the initial phase, so limiting the number of subsequent substitutions. The second experiment aimed at testing template updating in gradually degrading settings. In order to stress the system even more, we allowed template deletion. This implies that we can have PTG<1 even for a fully operational agent. We took again T_4 as the agent to be affected, but updating might have involved all agents at the same time. Our choice is motivated by the higher readability of results. We introduced a strong element of difficulty, since the initial gallery of T_4 included images with occlusions from set 8, while the probe set for all agents was the union of sets 2 and 3. This implied a higher number of substitutions or deletions followed by re-insertions, which increased the time needed to reach a stable configuration (Figure 2(b)) with respect to the previous experiment. We also noticed that more than 90% images with sun glasses (the initial gallery) were substituted, and this agrees with the better identifiability of images from sets 2 and 3 (smiling and angry) with respect to occluded images. The last experiment was the worst case (Figure 2(c)): we had a pre-existing gallery (set 6), while the probe also included images from set 8. Nevertheless, the algorithm showed the ability to select, among the different probes, the useful ones to be substituted in the gallery for T_4, so obtaining a substantial improvement of performances. The value for PTG is lower than in the preceding experiments, due to the higher number of responses labelled as unreliable by the NCT system composed by T_1, T_2, and T_3.

Besides accuracy, we also observed the frequency of the different operations (insertion, substitution, deletion) during gallery creation. The algorithm alternated phases with a significant amount of creations of new templates, and phases with more frequent updates. In the first case the gallery was being populated, so that there was an increase of G_T/G, but this also corresponded to an increase of EER and to a reduction of RR (i.e., to a temporary reduction in accuracy). However, during the second kind of phase, though preserving a good G_T/G, EER decreased and RR increased (performances improved). On the other hand, deletions were more infrequent. When the agent reached stability, the number of operations became negligible.

Figure 2(c) shows that the system converges more slowly in a critical situation, however the reached state is satisfactory enough, since, using set 6 as probe, it presents an RR of 80% compared with the 87% value in the optimal case (set 1, i.e. neutral, as gallery and set 6 as probe 2).

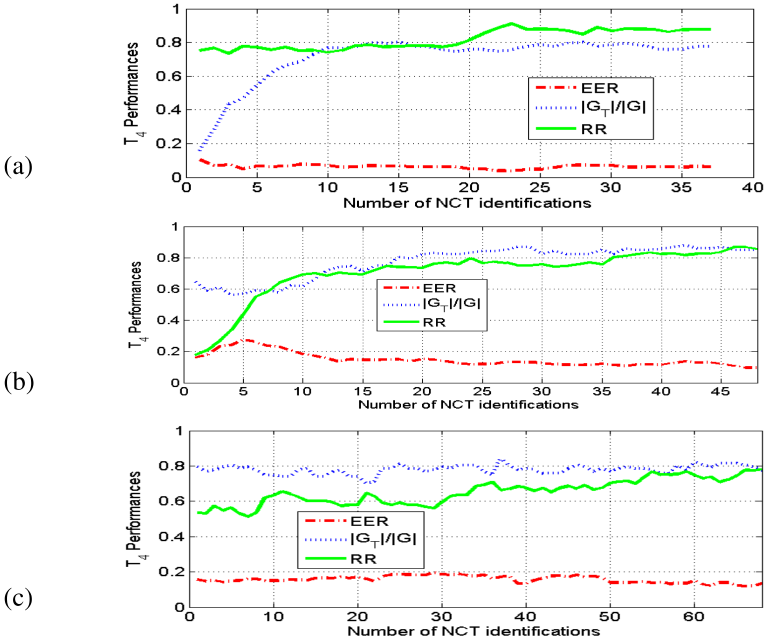


Fig. 2. Performance of the algorithm with different starting galleries: a) empty; b) images with occlusions; c) images with right light and occluded probes

5 Conclusions

We presented an algorithm for unsupervised template labeling for an agent, when it enters a fully operational multiagent architecture. We showed its ability to select significant templates even in case of template updating, and in non optimal conditions.

References

1. De Marsico, M., Nappi, M., Riccio, D., Tortora, G.: A multiexpert collaborative biometric system for people identification. JVLC 20(2), 91–100 (2009)
2. De Marsico, M., Nappi, M., Riccio, D.: FARO: FAce Recognition Against Occlusions and Expression Variations. IEEE Trans. On Systems, Man and Cybernetics–Part A 40(1), 121–132 (2010)

3. De Marsico, M., Nappi, M., Riccio, D., Tortora, G.: NABS: Novel Approaches for Biometric Systems. Accepted for Publication. In: IEEE Trans. on Systems, Man, and Cybernetics—Part C (available online)
4. Fatukasi, O., Kittler, J., Poh, N.: Estimation of Missing Values in Multimodal Biometric Fusion. In: 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems, BTAS 2008, Arlington, VA, pp. 1–6 (2008)
5. Martinez, A.: Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. IEEE Trans. on PAMI 24(6), 748–763 (2002)
6. Rattani, A., Freni, B., Marcialis, G.L., Roli, F.: Template Update Methods in Adaptive Biometric Systems: A Critical Review. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 847–856. Springer, Heidelberg (2009)
7. Roli, F., Didaci, L., Marcialis, G.: Adaptive Biometric Systems That Can Improve with Use. In: Ratha, N.K., Govindaraju, V. (eds.) Advances in Biometrics - Sensors, Algorithms and Systems. Springer, Heidelberg (2008)
8. Uludag, U., Ross, A., Jain, A.: Biometric template selection and update: a case study in fingerprints. Pattern Recognition 37, 1533–1542 (2004)

Automatic Bus Line Number Localization and Recognition on Mobile Phones—A Computer Vision Aid for the Visually Impaired

Claudio Guida, Dario Comanducci, and Carlo Colombo

Dipartimento di Sistemi e Informatica,
Via Santa Marta 3, I-50139, Firenze, Italy
{guida,comandu,colombo}@dsi.unifi.it

Abstract. In this paper, machine learning and geometric computer vision are combined for the purpose of automatic reading bus line numbers with a smart phone. This can prove very useful to improve the autonomy of visually impaired people in urban scenarios. The problem is a challenging one, since standard geometric image matching methods fail due to the abundance of distractors, occlusions, illumination changes, highlights and specularities, shadows, and perspective distortions. The problem is solved by locating the main geometric entities of the bus façade through a cascade of classifiers, and then refining the matching with robust geometric matching. The method works in real time and, as experimental results show, has a good performance in terms of recognition rate and reliability.

Keywords: Visual machine learning, object recognition, geometric methods, accessibility software.

1 Introduction

In the last few years, advanced technologies have greatly contributed to improve the mobility and autonomy of disabled people. For example, new interaction paradigms based on voice commands or eye movements have recently been developed for people with motor disabilities, and used to control wheeled chairs, or to communicate with the rest of the world through a computer.

Computer vision is one of the most important and useful technologies for the visually impaired (blind and low-sighted) people both in outdoor and indoor scenarios. The basic idea is that cameras can be used as additional eyes, whose images are automatically analyzed by the software so as to support the visually impaired in their everyday tasks. Most of the current work is focused on the development of specific methods for scene analysis/enhancement, and the construction of special devices that blind people can bring with themselves. In [5], a computer vision tool for producing automatic descriptions of video material is presented. Such descriptions can be useful to let blind users to follow better their favorite TV programs, especially for the parts containing little dialogue.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 323–332, 2011.

In [7], the authors describe a specific head-mounted device that can scan interesting parts of the scene and retrieve useful information for blind users. In [2], a special image contrast enhancement method is described, that dramatically improves the fruition of photos, text and other visual material by low-sighted people. The system described in [1] can be used to localize and recognize text in urban scenes; the system employs a portable computer placed in a back-sack, and one camera that is mounted on the user's shoulder.

In most urban scenarios, visually impaired people experiment everyday the difficulty of getting into the right bus. This is because the bus line number is typically provided only in a visual way to the people at the bus stop. In this paper, we propose an innovative method for localizing and recognizing in a completely automatic way the oncoming bus line number. Since the method uses only a standard smart phone (both for image grabbing and processing), no specific device is required. Image analysis is carried out by explicitly taking into account all the challenges arising in a real outdoor context, such as the illumination changes, highlights due to specular surfaces, the presence of occlusions, distractors, and perspective deformations. The method employs a careful combination of geometric and machine learning-based computer vision techniques, by which a satisfactory recognition performance is achieved.

The paper is organized as follows. In Section 2, an overview of the approach is provided, and then each of its main computational phases are described in detail. In Section 3 experimental results are presented. Finally, conclusions and future work are addressed in Section 4.

2 Our Method

Our method allows a visually impaired person standing at the bus stop with his mobile phone in hand to know the line number of the oncoming bus. As the bus is heard, the user directs the phone camera towards it, starting the acquisition from the mobile camera. The method can recover the line bus number from a single image of the bus. However, in order to attain a higher recognition reliability, several consecutive live images of the acquired sequence are processed separately, and the most frequent result is returned to the user via voice synthesis.

The method is split into several computational modules (Fig. 1).

First of all, the oncoming bus is localized inside the image and only the image region of interest including it ("Bus ROI") is considered for further processing ("Bus localization" module). As a result of this selection, unnecessary visual information is excluded from the next processing stages, thus speeding up computations. To further speed up the bus localization process, the original image (3 Mpixels) is reduced to a size of 0.75 Mpixels. The localization module exploits trained classifiers; it will be explained in detail in Section 2.1.

Once the bus ROI has been found, a further image cropping is performed ("Line number localization" module), in order to isolate the line number region ("Number ROI"). To this aim, a template description of the bus façade is matched with the image content of the Bus ROI (see Section 2.2).

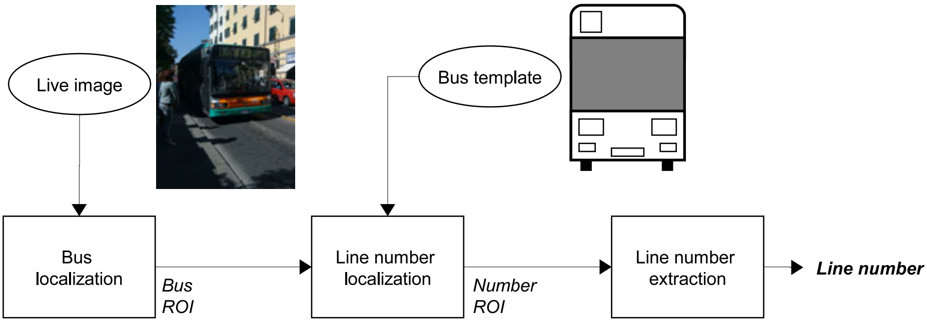


Fig. 1. Flow-chart of the proposed method

The last stage of our method is dedicated to reading the number of the bus line ("Line number extraction" module), discussed in Section 2.3. For this task, we first binarize the Number ROI, and then use trained classifiers to read the bus line number from the binarized subimage.

2.1 Bus localization

To detect the bus presence and to localize it in the image we use the machine learning algorithm proposed by Viola and Jones in [10]. Originally proposed for face recognition, the algorithm exploits a cascade of weak classifiers and a boosting training method [4]. In our case the classifiers were trained for bus recognition.

Fig. 2 shows three examples of bus localization, related to photos taken from a mobile phone under different viewing/lighting conditions and for different bus models. The rectangle delimits the detected bus ROI. Notice that the bus is correctly localized even in the presence of occlusions (Fig. 2(a)), and in the presence of distractors as the van in Fig. 2(c). Distinguishing between real buses and van or trucks is a very hard task for blind people, that rely only on acoustic cues.

To train the classifier, we took several photos at bus stops as positive examples to feed the boosting training algorithm. To enforce robustness to pose and light variation, we also created 100 virtual views from each original image. We trained the model using about 1500 positive examples and 2100 negatives.

2.2 Line Number Localization

Once the bus is localized, all subsequent processing is limited only to the bus ROI. In order to make the method robust to pose variations, we use a template bus façade description to geometrically rectify the projective distortions arising as the result of image projection. Other image distortions that make line number localization challenging arise with variations in light and reflection phenomena.

Indeed, classic matching techniques with a frontal image of the bus façade are not useful in this context, due to the presence of important specularities, that

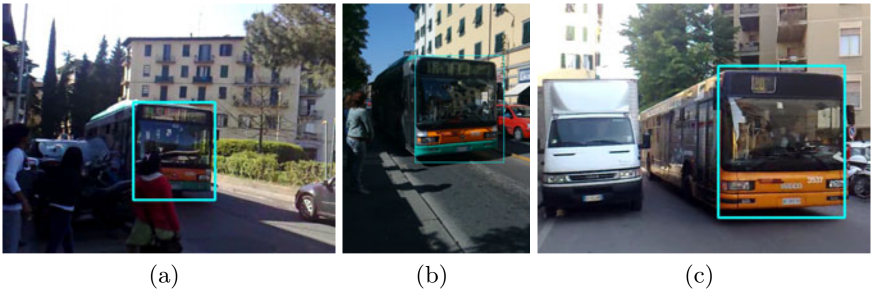


Fig. 2. Viola & Jones bus localization: The classifier is run on three images of an urban scene. (a): Notwithstanding people occlusion, the bus is correctly localized. (b): Correct localization under different lighting conditions. (c): Correct localization of a different bus model. **Best viewed in color.**

make standard rectification methods prohibitive, due to the difficulty to extract robust features and match their descriptors. Figs. 3(a) shows an example where several wrong SIFT matches occur between a frontal view of a bus facade and a bus ROI image, thus making the rectification totally incorrect (Fig. 3(b)) even with the use of robust estimation techniques such as RANSAC [6] or LMedS [11]. This is because the number of outliers in the matching set is simply too high.

The template description of the bus façade allows us to overcome most difficulties: In Fig. 3(c) a good rectification of the façade, by exploiting our template, is shown.

The template was built by selecting some geometrically distinctive features of the façade (see Fig. 4(a)): Some elements of the bottom part (the four lights, the plate and the central logo) and the top line of the bus façade. The template description contains also the line number ROI (the rectangle in the top left region of the bus façade), to be used in the next step. For each of the template features

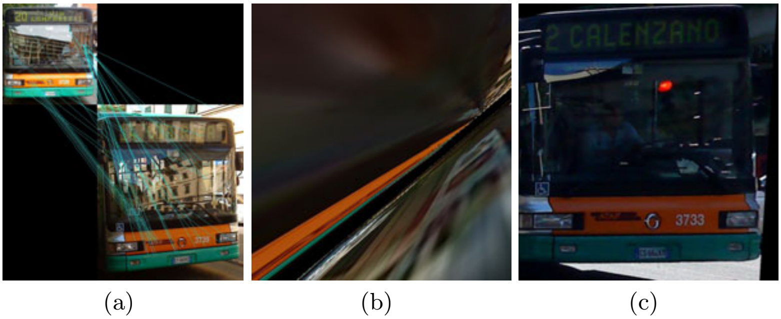
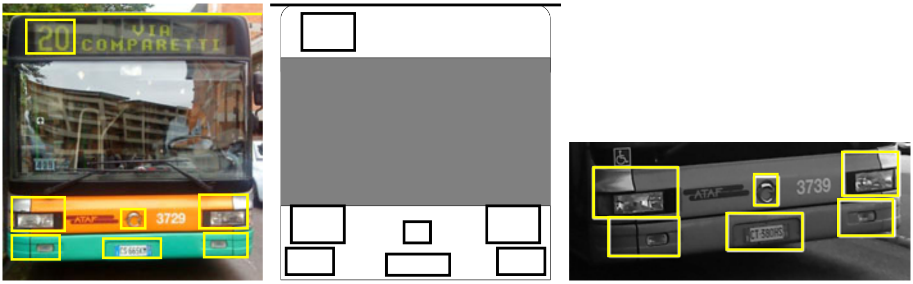


Fig. 3. Example of wrong rectification. (a): Fake matches. The upper-left part of the picture shows the image with bus façade frontal view, while in the bottom-right part the bus ROI to be matched is reported. (b): Wrong rectification of the bus ROI. (c): A correct rectification of the bus ROI by using the template. **Best viewed in color.**



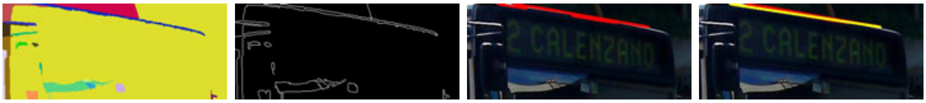
(a) (b) (c)

Fig. 4. (a): Annotation of the distinctive features of the bus façade. (b): Template model; the annotated elements with their position are shown with bold lines. (c): The distinctive elements found by the classifiers $\mathcal{C}_i$.s **Best viewed in color.**

the corresponding position within a normalized frontal view of the façade was annotated (Fig. 4(b)), and a separate classifier $\mathcal{C}_i$ was trained (with the exception of the top line and bus number area). The annotated positions are exploited to recover the rectifying homography of the bus façade.

Each classifier $\mathcal{C}_i$ applied to the bus ROI of the live image returns the minimum enclosing rectangle of the detected object from which its center $\mathbf{x}_i$ is calculated, and associated with its corresponding point $\mathbf{x}'_i$ in the template. Fig. 4(c) shows a result of the six classifiers for the bottom part of the template.

To retrieve the upper border line in the bus ROI we use classic image analysis methods such as Canny edge detection and Hough transform [9]. High noise present in the upper side of the image, i.e. sky, trees, buildings, etc. makes line retrieval quite difficult. To mitigate this problem, we apply the graph-based segmentation discussed in [3]. Fig. 5 shows an example of intermediate image processing which leads to upper border line retrieval. The retrieved line $\mathbf{l}$ is put into correspondence with the top border line $\mathbf{l}'$ of the bus template.



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

Fig. 5. Retrieving the upper line. (a): Result of graph-based segmentation. (b): Result of Canny edge detector. (c): Filtering of shorter lines. (d): Approximate upper border retrieved. **Best viewed in color.**

The rectifying homography $\mathtt{H}$ of the bus façade is estimated by combining into a linear Least Squares problem the constraints [6]

$$\mathbf{l}' = \mathtt{H}^{-\top}\,\mathbf{l} \tag{1}$$

$$\mathbf{x}'_i = \mathtt{H}\,\mathbf{x}_i \quad i = 1 \ldots 6. \tag{2}$$

2.3 Reading the Number

Once the rectified bus ROI is computed, we can safely localize the number ROI by simply selecting the corresponding area from the template, and mapping it onto the rectified ROI. Figs. 6(a) and 6(d) illustrate two examples of number ROI localization.

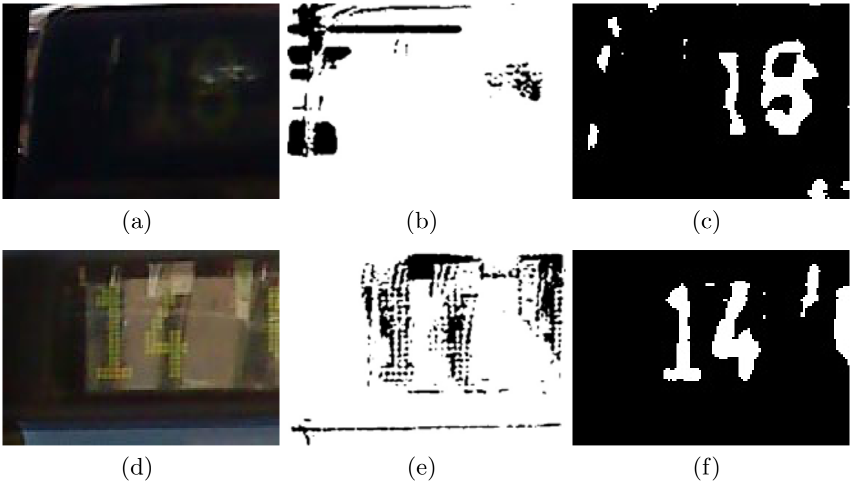
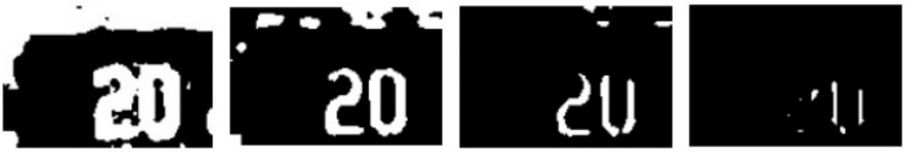

(a) (b) (c)
(d) (e) (f)

Fig. 6. (a, d): Shadows and specularities make reading the line number a difficult task. (b, e): standard OCR binarization techniques fail the extraction of the bus line number. (c, f): The binarization obtained with our method. **Best viewed in color.**

To perform line number reading, it is important that the contents of the number ROI are properly binarized. Binarization is also a common and crucial procedure in OCR techniques. As one can see from Figs. 6(b) and 6(e), the severe shadowing and mirroring phenomena make classic OCR binarization methods as [8] unuseful for the bus line number reading task.

We propose here an alternative and original binarization method based on complementary colors, that exploits the knowledge of the bus number color. Working in the HSV color space, we can easily find the complementary of the bus number color, and make an adaptive thresholding operation to filter all colors that don't have the same hue. Figs. 6(c) and 6(f) show the results of this new binarization method on the example illustrated in Figs. 6(a) and 6(d). Notice the dramatic improvement in image quality after binarization.

To make the process invariant to light variations, we also developed an adaptive thresholding method. The process tries to find the optimal thresholding value τ by iteratively binarizing the original image, until the number of white points in the entire image is inside a predefined interval. Fig. 7 shows an



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

Fig. 7. Some results of our binarization method obtained by varying the threshold τ. (a) $\tau = 20$, (b) $\tau = 50$, (c) $\tau = 70$ and (d) $\tau = 90$. Best thresholding value τ is 50.

example of some iterations of our binarization algorithm. We can see that there is a particular thresholding τ value for which the number is quite clear. In this particular case $\tau = 50$.

Once the binarization operation is completed, the bus line number can be extracted. Once again machine learning is exploited for this task. In particular, we trained a classifier for each line number (not digit) to be recognized. The process of classification is as follows. Classifiers are put in a chain of ascending precision of classification, i.e., the first classifiers in the chain have better performance. We use then a first-win-takes-all schema to recognize the number.

3 Experimental Results

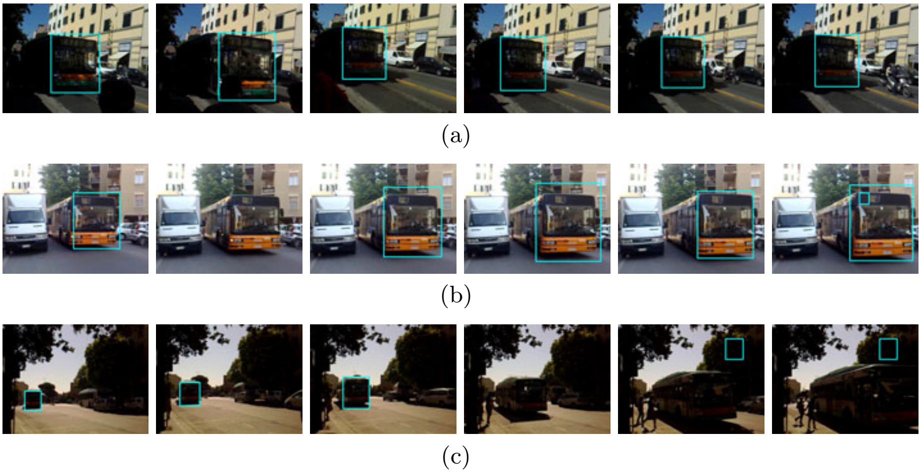
In this section we will show experimental results obtained by applying our method to a dataset collected on the road.

The performance of the bus localization module is first addressed, and then results on the recognition of the bus line number are provided.

3.1 Bus Detection and Localization

To test the classifier for the bus localization task, we used various video sequences captured from mobile phones. Fig. 8 shows some results from three particular sequences. As one can see from Fig. 8(c), not all the frames of the sequence return good results. This is not a critical problem, because our method works on multiple images and the line number detection is activated as soon as a bus is recognized. For the same reason, false positive results are not a severe problem.

Table 1 illustrates detailed statistics concerning all the frames of the sequences. Bus detection rate is quite high in most of the sequences, with the exception of the first, which has just 6 frames over 32 with correct bus localization. Some frames of this sequence are indeed reported in Fig. 8(c): the very dark appearance of the bus façade, due to the sun in front of the camera, is the main reason of the low detection rate for that sequence. Although the low bus localization rate, our method successfully recovers the correct line number in all the 6 frames where the bus is detected.


(a)

(b)

(c)

Fig. 8. Example of bus localization in three video sequences. **Best viewed in color.**

Table 1. Results of bus detection for each video sequence

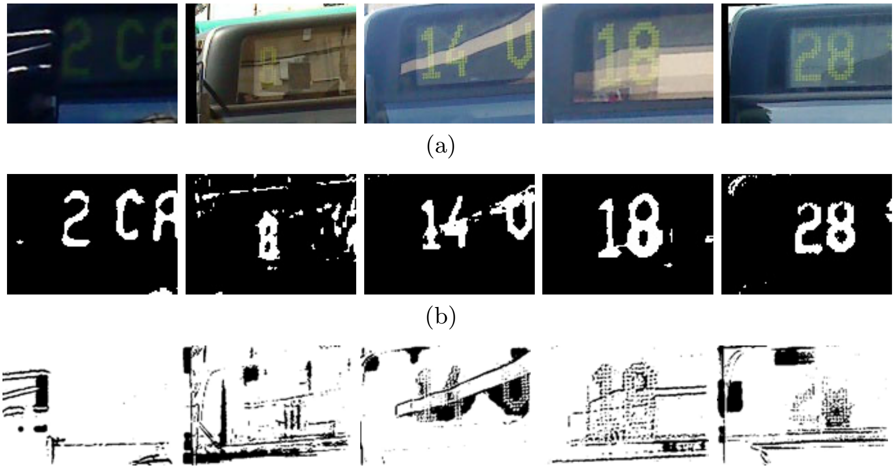
Sequence	Frames	Detection	False Positive	Rate
No. 1	32	6	2	18%
No. 2	40	40	n/a	100%
No. 3	17	11	1	64.7%
No. 4	24	23	n/a	95.8%
No. 5	46	33	3	72%
No. 6	48	27	1	56.2%

3.2 Recognition of Bus Line Number

In this section the performance of our method for the bus line number recognition task is presented. A dataset composed by 50 images taken from frames not previously used in the training session (both for the bus localization and for the detection of the objects of interest) was exploited; in particular 10 images for 5 bus lines (2, 8, 14, 18 and 28) were used.

For each bus line considered in the dataset, an example of the rectified number ROI is provided in Fig. 9(a), while in Fig. 9(b) the corresponding binarized images are shown. Fig. 9(c) shows the binarization results obtained with the binarization technique of [8]. Working on the complementary color space makes our binarization algorithm invariant to specularities, shadows and changes in illuminations, while [8] fails.

Table 2 reports the detection rates on our dataset and shows that, with the exception of classifier relative to the line number 8, every bus line number is correctly recognized with 100% rate. The classifiers in the first-takes-all chain are ordered as 14 (the strongest), 18, 28, 8, 2 (the weakest). We found that several mutual false positives occur between the classifiers for the bus line numbers



(a)

(b)

(c)

Fig. 9. Some results obtained applying our proposed binarization method. (a): The rectified number ROI. (b): our binarization results. (c): the binarized image obtained with [8]. **Best viewed in color.**

Table 2. Results of our rectification and binarization algorithm to the entire dataset

Line Number	Detection rate
2	100%
8	75%
14	100%
18	100%
28	100%

2 and 8, because of the particular font used by the transport company, and probably this is the reason of the weak performance of the classifier for the bus line number 8.

4 Conclusions and Future Work

In this paper, an auxilium for visually impaired people aimed at locating and recognizing an incoming bus line number was presented. The method combines geometric computer vision with machine learning so as to achieve robustness with respect to highlights, specularities, shadows, occlusions, and so on. Experimental results show that the method has a higher reliability with respect to traditional geometric matching methods and standard OCR techniques. Future work will address using multiple bus templates, and dealing with several oncoming buses simultaneously. To achieve these goals we need to modify our detection and localization algorithm, in order to speed up the overall process. Bus localization will be performed within a pyramidal framework, and faster alternatives to the Viola and Jones approach will be investigated for the detection of template elements and the localization of the number region.

References

1. Chen, X., Yuille, A.: A time-efficient cascade for real-time object detection: with applications for the visually impaired. In: Proc. Conf. Computer Vision and Pattern Recognition (2005)
2. Choudhury, A., Medioni, G.: Color contrast enhancement for visually impaired people. In: Proc. 3rd Computer Vision Applications for the Visually Impaired (2010)
3. Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. International Journal of Computer Vision 59, 167–181 (2004)
4. Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Proc. 2nd European Conference on Computational Learning Theory, pp. 23–37 (1995)
5. Gagnon, L., Chapdelaine, C., Byrns, D., Foucher, S., Hritier, M., Gupta, V.: Computer-vision-assisted system for videodescription scripting. In: Proc. 3rd Computer Vision Applications for the Visually Impaired (2010)
6. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge (2004)
7. Pradeep, V., Medioni, G., Weiland, J.: Robot vision for the visually impaired. In: Proc. 3rd Computer Vision Applications for the Visually Impaired (2010)
8. Seeger, M., Dance, C.: Binarizing camera images for ocr. In: Proc. 6th International Conference on Document Analysis and Recognition, pp. 54–58 (2001)
9. Trucco, E., Verri, A.: Introductory Techniques for 3-D Computer Vision. Prentice Hall, Englewood Cliffs (1998)
10. Viola, P., Jones, M.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004)
11. Zhang, Z.: Parameter estimation techniques: A tutorial with application to conic fitting. Image and Vision Computing 15(1), 59–76 (1997)

The Use of High-Pass Filters and the Inpainting Method to Clouds Removal and Their Impact on Satellite Images Classification

Ana Carolina Siravenha, Danilo Sousa, Aline Bispo, and Evaldo Pelaes[⋆]

Signal Processing Laboratory, Federal University of Para (UFPA), Belem, PA, Brazil
{siravenha,danilofrazao,pelaes}@ufpa.br, aline.bispo@itec.ufpa.br

Abstract. This paper proposes a new technique to smooth undesirable elements of the atmosphere, such as fogs, clouds and shadows, which damage and lead to loss of image data. In our approach, an efficient way to detect clouds and shadows is presented. The method applies constants related to such undesirable elements, as well as a High boost Filter in the homomorphic filtering for scattered clouds removal. We highlight the use of the Inpainting method, which replaces contaminated pixels using a nearest neighbor interpolation. Beside this, the proposed algorithm adopts a morphologic opening of the image that aims to suppress some isolated occurrences in the scene. The results are evaluated by Kappa coefficient and PSNR index, proving the good performance of the method.

Keywords: Cloud removal, High boost filtering, cloud detection, inpainting.

1 Introduction

One of the main problems related to remote sensing of images, whether aerial or satellite, is the presence of undesirable elements in the atmosphere such as fog, mist, clouds and the shadows from them. Since the sensing becomes a constant target of these atmospheric components, which commonly occurs in tropical areas, activities such as environmental or urban monitoring, or any other study for extraction of relevant information to users, become impaired.

Due to these problems, various techniques have been developed for the removal or, at least, the mitigation of these effects, especially the clouds, which are more frequent in this type of image. According to the literature, the nature of these techniques depends on the kind of cloud, dense or sparse. The removal of dense clouds and their shadows are basically divided into two approaches, the first relates to the use of a non-cloud cover reference image in a multi-temporal analysis [18], or even doing an interpolation with contribution of a radar image (SAR) [8]. The other approach attempts to estimate what is being covered

[⋆] This work was supported by *Fundação de Amparo à Pesquisa do Estado do Pará* and Federal University of Para.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 333–342, 2011.

by cloud and/or shadow in image, using the same or similar approaches to the inpainting method [2,12,14].

On the other hand, when it comes to the removal of scattered clouds, the most used method is the homomorphic filtering [3,16,13]. In [7], an evaluation was made among five High-pass filters (HPF), highlighting one that gives the best results, the Butterworth filter.

This paper proposes a technique for removing the elements of the atmosphere that damage and lead to loss of image data. Our method detects clouds and shadows with the aid of constants related to such elements, as well as a High-Boost Filter (HBF) [10,17] to the filtering of scattered clouds. In a comparison with the approach applied in [7], the proposed method was found to be more efficient. The HBF does not totally eliminate the low frequency components of the image, thereby causing smaller information losses. Another important feature of this approach is the use of inpainting method, generally employed for the removal of large objects in an image [1,2]. In this paper, the impainting method is applied for the redefinition of dense blocks of clouds and shadows. By executing a nearest neighbor interpolation [9,6] and using only the information of image pixels to be processed, this method is made more versatile and adaptive to the context. At last, a morphological operation of opening of the image is also applied in this proposed method.

The paper is divided as follows. Section 2 describes the theory of all techniques used in this work as well as how the algorithm works; section 3 shows the results after the application of the algorithms and the evaluation of the methods; and finally, section 4 shows the conclusions about the described method.

2 Methodology

In this section we present the basic techniques employed to develop our cloud and shadow removal algorithm. Each technique is presented and its contribution to the proposed method are described.

2.1 Cloud and Shadow Detection

The basis of the shadows and clouds detection is based on [7], in which is made a separation of regions with different characteristics in the image, in order to improve results. This division is made considering statistical measures of the image, detecting dense and scattered clouds areas.

The presence of clouds in a remote sensing image is usually associated with the presence of shadows. The Sun angle at the scene capture and/or the scene capture in an off-nadir angle[1] are some explanation for the shadows formation in remote sensory images. For this purpose, it was added to the cloud detection algorithm, the shadows detection capability.

[1] The term *off-Nadir* refers to non-orthogonal imaging between the sensor and the imaged object.

The difference on the lighting conditions, on the sensor characteristics and on the surface, make that the images have different ranges of gray levels representing a class. For example, under opposite lighting conditions a same region can be easily labeled as dense and non dense vegetation, or a dense vegetation region can be mislabeled as a shadow region. In order to improve the technique described in [7], making the regions separation more flexible, we added two constants to the detection algorithm, called cc and sc, cloud and shade constants, respectively. Both constants have default value equals to 1, and after the tests, was noted that the range from 0 to 3 returns optimized results. Therefore, by the change of these values we can differentiate more precisely the classes of a scene.

The process of regions separation is expressed by

$$f(x,y) = \begin{cases} i(x,y) < (sc \times i_{m-dp}), & i(x,y) \epsilon\ 0; \\ (sc \times i_{m-dp}) < i(x,y) < i_m, & i(x,y) \epsilon\ 1; \\ i_m < i(x,y) < (cc \times i_{m+dp}), & i(x,y) \epsilon\ 2; \\ i(x,y) > (cc \times i_{m+dp}), & i(x,y) \epsilon\ 3. \end{cases} \tag{1}$$

where $i(x,y)$ is equal to the pixel value of the noise image (cloud or shadow), i_m represents the average value of the noisy image, $i_{m+dp} = i_m + \sigma_i$ and $i_{m-dp} = i_m - \sigma_i$. Thus, regions defined as 0 are the shadow regions of the image, those defined as 1 are free from any kind of noise, while the regions defined as 2 are labeled as containing scattered clouds and finally those defined as 3 are detected as dense clouds.

2.2 Image Opening

The morphological filters refer to the study of geometric structure of the entities present in an image. Being the filters of dilation and erosion basic morphological operations and based on set theory, this technique involves the interaction between an image A (the object of interest) and a structuring element B. In general, most of the morphological operations are based on simple operations of expansion and shrinkage [5].

Opening generally smooths the contours of an image, breaking narrow isthmus and eliminating thin protrusions, and is defined as follows: the opening of A by B is given by the erosion of A by B, followed by dilation by B, that is, $A \circ B = (A \ominus B) \oplus B$.

Therefore, after opening the image, small objects inside a larger tend to be extinct. Our interest in applying such a morphological transformation, relies on the fact of small objects contained in large blocks defined as noise can worsen the interpolation phase, during the inpainting method. Given this, the opening operation can eliminate this problem, generating a better redefinition of image pixels damaged by atmospheric action.

2.3 Homomorphic Filter

The images usually consist of light reflected from objects. The basic nature of an image can be characterized by two components: (1) the amount of light coming from the incident source on the scene and (2) the amount of reflected light by objects in the scene. These bands of light are called components of *illumination* and *reflection*, and are denoted by $l(x, y)$ and $r(x, y)$, respectively. The functions l and r combine multiplicatively to produce the F image:

$$F(x, y) = l(x, y)r(x, y), \tag{2}$$

where $0 < l(x, y) < \infty$ e $0 < r(x, y) < 1$.

The Fourier transform of the product of two functions is not separable [7] and it is desirable to manipulate the image in frequency domain, then we apply the natural logarithm function that approximates the function $F(x, y)$ to the form of a sum of the components l and r

$$z(x, y) = \ln F(x, y) = \ln l(x, y) + \ln r(x, y). \tag{3}$$

In frequency domain, we can manipulate the image in terms of low and high components, or illumination and reflection, separately. Then, we apply a Fast Fourier Transform on $z(x, y)$, the natural logarithm of $F(x, y)$

$$\mathscr{F}(z(x, y)) = \mathscr{F}(\ln F(x, y)) = \mathscr{F}(\ln l(x, y)) + \mathscr{F}(\ln r(x, y)). \tag{4}$$

Therefore, considering that Z, L and R are Fourier Transforms of z, $\ln l$ and $\ln r$, respectively, then

$$Z(w, v) = L(w, v) + R(w, v). \tag{5}$$

The advantage of the use of high pass High-boost filter to the cloud removal applications is that it does not completely suppress the lower frequencies. Thus, the filter $H(.)$ is applied to the function Z and assumes

$$S(w, v) = H(w, v)Z(w, v) = H(w, v)L(w, v) + H(w, v)R(w, v) \tag{6}$$

We can take the inverse Fourier transform of Eq. 6

$$s(x, y) = \mathscr{F}^{-1}(H(w, v)L(w, v)) + \mathscr{F}^{-1}(H(w, v)R(w, v)), \tag{7}$$

and finally, as z was obtained using the logarithm of the original image F, the reverse process produces the desired image $\hat{F}$

$$\hat{F} = \exp s(x, y) = \exp\left(l'(x, y)\right)\exp\left(r'(x, y)\right) = l_0(x, y)r_0(x, y). \tag{8}$$

By applying a multiplicative factor of amplification (A) before subtraction of the low frequencies (LPF), we obtain a filter HPF High boost. Thus,

$$\begin{gathered} Highboost = (A)(original) - LPF \\ Highboost = (A - 1)(Original) + HPF \end{gathered} \tag{9}$$

where $HPF = Original - LPF$. If $A = 1$, we have a simple filter high-pass. When $A > 1$, a part of the original image is maintained in the output. A Butterworth HPF used in the formulation of this proposed High boost filtering.

Homomorphic filtering using High boost is applied only in the areas labeled as 2, i.e., areas where there is presence of scattered clouds, mists and fogs. The filtering in these regions is suitable, because the region of cloud is a low frequency region due to the homogeneity of its pixels.

Thus, the homomorphic filtering process can be summarized by the Fig. 1.

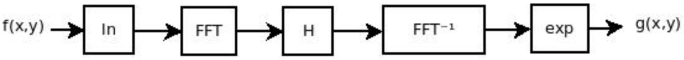


Fig. 1. Homomorphic Filter Schema [3,16]

2.4 Inpainting

Areas labeled as 3 during the separation of regions (Subsection 2.1) are the thickest clouds present in the images, which suffer very little or even no effect of homomorphic filtering. The treatment of these areas is done by a method called inpainting. This method aims to remove large objects, fill gaps, undefined or damaged regions of an image, in order to restore or make it more visible [2,12]. In this paper, we consider as damaged regions to be treated, areas of dense clouds, as done in [14], and shadow regions.

The Inpainting method uses the technique of nearest neighbor interpolation and the effect created is that the nearest pixel appears larger, because its value is assigned to an undefined pixel [9].

Taking a point x to be interpolated and various points x_k neighbors to such, which belong to a set of values of samples of image pixels f_k, a calculation of Euclidean distance is made between this value to be estimated and the neighboring sample values, according to the following equation: $||x - x_k||$. The value of x_k which yields the lowest result in the last equation, has its intensity value assigned to the pixel x, and so continues until all undefined pixels are filled [6].

2.5 Algorithm of the Proposed Method

The steps of the proposed algorithm are shown in sequence:

1. The regions of the image F are mapped as described in Subsection 2.1. The result of this step is an mapped image among the four categories described, called FM.
2. The FM mapped image passes through an opening operation, generating a new image FMO, which will be the base for mapping in the fusions.
3. The *pixels* of image areas F labeled as 0 and 3 are treated as undefined, then the *inpainting* method (Subsection 2.4) assigns new values to these pixels, generating an interpolated image FI.
4. The images F and FI are joined to form the fused image FF. This fusion is done by the combination of the areas 0 and 3 of the FI image, with areas 1 and 2 of the image F.

5. The original image F is passed through homomorphic filter (according Subsection 2.3), removing scattered clouds, and producing an image called FH.
6. The last fusion of images to form the resulting image (FR) is made between images FF and FH. From image FH are extracted the pixels previously classified as type 2, and the remaining pixels are got from the image FF.

The flowchart of the process described above is shown in the diagram of Fig. 2:

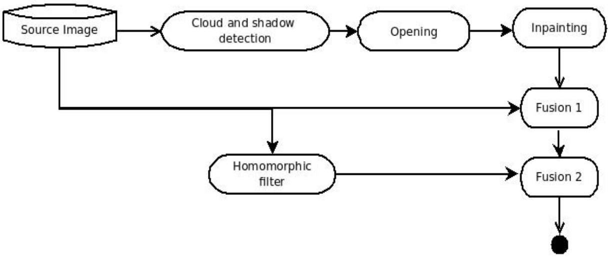


Fig. 2. Flowchart of the proposed algorithm

3 Results and Discussions

In order to exemplify the procedure described previously, we take as *Original image*, we have a scene of city of Rome (Italy), captured by the WORLDVIEW-2 satellite sensor, in 2009. The scene is contaminated with three types of atmospheric damage mentioned in this paper: scattered clouds, dense clouds and shadows (4% of total scene area).

The steps to remove these elements will be evaluated according to the Kappa coefficient of the image [7] and the peak signal to noise ratio (PSNR) index [15,4], based on similarity and compatibility of information between the original and reference images, where bigger values represent the best index achievable. We use two reference images for such evaluations, which passed by a process of manual removal of atmospheric noise simulating optimal pictures, one free of scattered clouds and other free of all kind of noise.

Firstly, the Butterworth filter used by [7] was tested in comparison with High boost approach. In order to remove scattered cloud, the original image was submitted to both filters and compared with a reference image free of scattered clouds. According the metrics cited before, the homomorphic filtering using High boost presented highest accuracy, with $Kappa = 0.7788$ and $PSNR = 21.63dB$, whilst the HPF Butterworth approach presented $Kappa = 0.3973$ and $PSNR = 19.82dB$, showed in Fig. 3. Rather than many HPF, the filter High boost does not totally eliminate the low frequency components, allowing them to assist in image interpretation. Thus, it becomes clear that the High boost filter is more efficient, producing visibly clearer results with less loss of information.

Aiming to remove dense clouds and shadows, still without the constant of regions detection, the inpainting method was applied to both high pass filtering

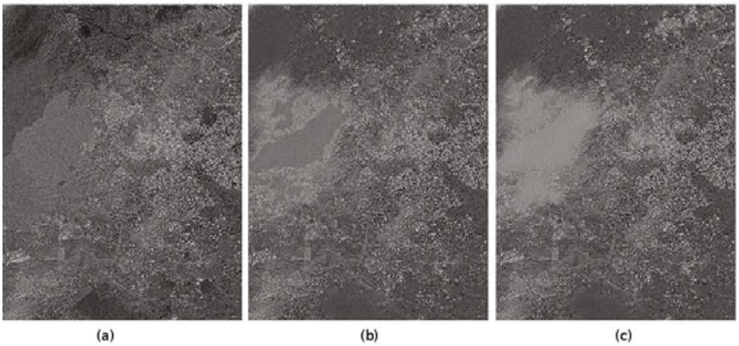


Fig. 3. First test: (a) Original image and scattered clouds removal results using HPF (b) Butterworth ($Kappa = 0.3973$ and $PSNR = +19.82dB$) and (c) High boost ($Kappa = 0.7788$ and $PSNR = +21.63dB$)

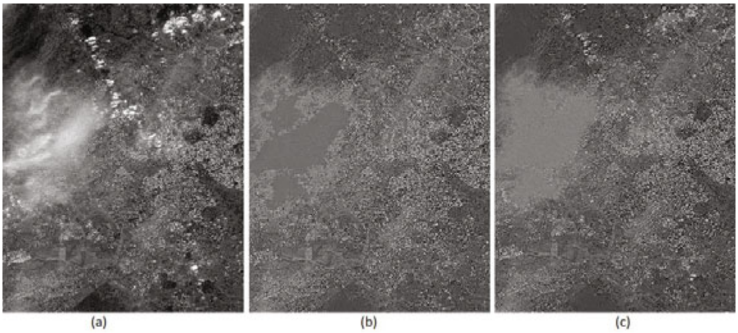


Fig. 4. Second test: (a)Original image and dense clouds and shadows removal results using HPF (b) Butterworth ($Kappa = 0.4618$ and $PSNR = +19.4dB$) and (c) High boost ($Kappa = 0.4737$ and $PSNR = +19.6dB$)

resultant images. To obtain the Kappa coefficient and PSNR indexes were used a manually produced free of noise image. Once more, the approach using High boost got better results, with $Kappa = 0.4737$ and $PSNR = 19.6dB$, over Butterworth approach, with $Kappa = 0.4618$ and $PSNR = 19.4dB$, showed in Fig. 4. Note that the approach that uses the HPF High boost, is slightly more efficient, even with the great similarity between the results. It is also visible the disappearance or smoothing of most noisy regions of the image, despite the decrease in Kappa and PSNR index values, since we have now a high degree of redefinition of pixels, due to the amount of dense clouds and shadows in original image.

Finally, by entering the constants cn and cs, for the detection of clouds and shadow, respectively, was compared the effect of this insertion in the same scene presented before. Once again, the Butterworth approach presented lower performance than High boost. The Butterworth indexes achieved were $Kappa = 0.5279$

and $PSNR = 19.64dB$, while the High boost indexes achieved were $Kappa = 0.5329$ and $PSNR = 19.94dB$. Therefore, besides the High boost filter remain slightly better than the Butterworth, we see a considerable improvement in efficiency of the method using both filters after use of constants(setted to $cn = 1.3$ and $cs = 0.8$). The results are showed Fig. 5.

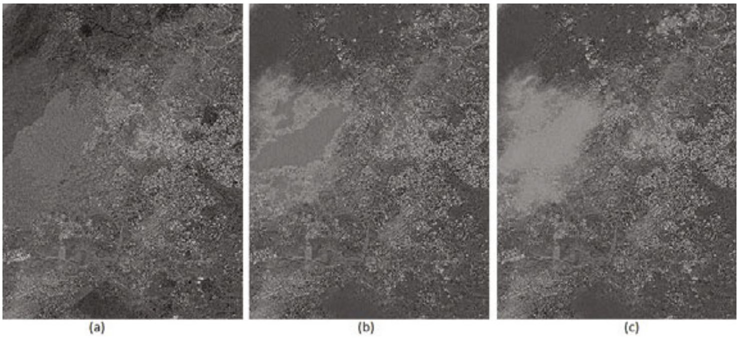


Fig. 5. Third test: (a) Reference image free of noise. The influence of cn and cs in clouds and shadows removal using HPF: (b) Butterworth ($Kappa = 0.5279$ and $PSNR = +19.64dB$) and (c) High boost ($Kappa = 0.5329$ and $PSNR = +19.94dB$)

Table 1. Evaluation of differents approaches in the proposed algorithm. The PSNR index is given in dB

Tested	Butterworth		High boost		Butterworth+constants		High boost+constants		Constants	
Images	Kappa	PSNR	Kappa	PSNR	Kappa	PSNR	Kappa	PSNR	cn	cs
Brasília	0.3940	18.54	0.5660	19.12	0.4626	18.41	0.6019	19.18	0.75	1.3
Belém 1	0.4471	24.44	0.7185	26.29	0.4572	24.57	0.7239	26.72	0.9	1.2
Fires	0.4669	21.33	0.4751	22.86	0.4858	24.49	0.5302	25.92	0.2	1
Alabama	0.5312	19.75	0.5368	20.23	0.6446	20.26	0.6702	21.16	1.5	1.1
Belém 2	0.6971	19.84	0.7035	20.61	0.7014	20.30	0.7492	20.94	0.9	1.2
Italy	0.6850	22.61	0.7131	23.04	0.7285	24.54	0.7295	24.76	0.5	1.25
Nasa	0.8217	26.78	0.8236	26.93	0.8351	26.26	0.8408	27.28	7	100
Ikonos 1	0.8312	24.23	0.8472	24.46	0.8445	24.43	0.8590	24.67	0.75	11
Lake	0.4982	12.86	0.5118	20.67	0.5452	13.82	0.6795	23.40	0.25	100
Ikonos 2	0.7814	24.79	0.7814	27.14	0.7862	26.65	0.7997	27.25	1.25	100
Andes	0.5292	17.89	0.5366	16.95	0.6025	18.58	0.6116	17.22	100	100
Capim	0.1803	17.19	0.6432	25.31	0.1867	17.42	0.6486	25.17	1.1	1.1
Macapá	0.2161	20.09	0.6487	26.13	0.7619	26.63	0.7501	27.36	0.8	1.5
Tucuruí	0.7583	21.91	0.8105	22.32	0.7991	24.31	0.8584	25.54	2.25	1.4
Belém 3	0.7378	22.66	0.8173	23.91	0.7710	23.04	0.8445	24.15	0.6	3

The method proposed was then applied to other 14 images. The Table 1 presents results in 15 different scenes (including the scene 'Italy' used to exemplify the method), where can be seen the superiority of the proposed method. It's important reinforce that such method, as well as others of the inpainting, works well for small damaged regions, however, when the reconstruction area is too large, its results can produces blurred and unreal areas, without information of texture [11,2]. This effect was evidenced on a dense cloud region of the scene shown in that article.

4 Conclusion

In this paper, we deal with the removal of undesirable elements in the atmosphere such as fog, mist, clouds and the shadows from them, that damage and lead to loss of image data. Thus, we propose a method for removing such elements, which uses several techniques. Was highlighted in our method, a more efficient way to detect clouds and shadows with the aid of constants related to such elements, as well as the use of High-Boost Filter during the homomorphic filtering, comparing with the approaches applied in [7]. By results, aided by Kappa and PSNR index, we conclude that HPF High boost is more efficient than Butterworth, and that the use of the constants for region detection improve the final results of the algorithm, leading to the disappearance or smoothing of most noisy regions, as showed in the resultants figures, including the Fig. 6.

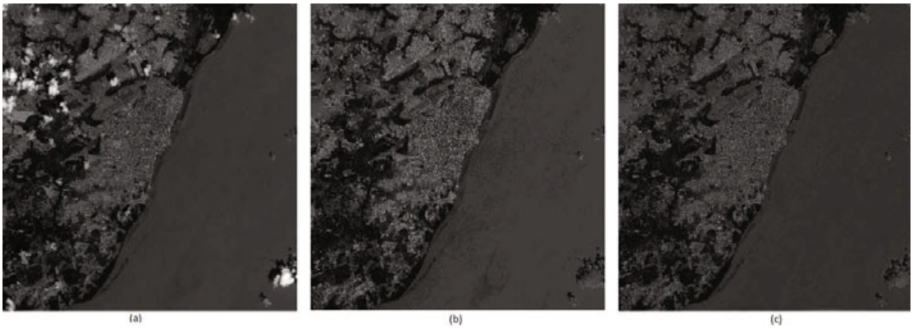


Fig. 6. Tested image 'Macapá': (a) Reference image free of noise. The influence of cn and cs in clouds and shadows removal using HPF: (b) Butterworth ($Kappa = 0.7619$ and $PSNR = +26.63dB$) and (c) High boost ($Kappa = 0.7501$ and $PSNR = +27.36dB$)

References

1. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, pp. 417–424 (2000)
2. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image Inpainting. IEEE Transactions On Image Processing 13, 1200–1212 (2004)
3. Delac, K., Grgic, M., Kos, T.: Sub-image Homomorphic filtering technique for improving facial identification under dificult illumination conditions. In: International Conference on Systems, Signals and Image Processing, pp. 95–98 (2006)

4. Delac, K., Mislav, G.: Handbook Of Data Compression. Springer, Heidelberg (2009)
5. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Addison-Wesley Publishing Company, Reading (2008)
6. Hale, D.: Image-guided blended neighbor interpolation of scattered data. In: 79th Annual International Meeting, Society of Exploration Geophysicists, vol. 28, pp. 1127–1131 (2009)
7. Hau, C.Y., Liu, C.H., Chou, T.Y., Yang, L.S.: The efficacy of semi-automatic classification result by using different cloud detection and diminution method. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2008)
8. Hoan, N.T., Tateishi, R.: Cloud removal of optical image using SAR data for ALOS applications. Experimenting on simulated ALOS data. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2008)
9. Htwe, A.N.: Image interpolation framework using non-adaptive approach and nl means. International Journal of Network and Mobile Technologies 1 (2010)
10. Kekre, H.B., Athawale, A., Halarnkar, P.N.: High payload using High Boost filtering in Kekre's Multiple LSB's algorithm. In: 2nd International Conference on Advances in Computer Vision and Information Technology (2009)
11. Kwok, T., Sheung, H., Wang, C.: Fast query for exemplar-based image completion. IP 19, 3106–3115 (2010)
12. Liu, H., Wang, W., Bi, X.: Study of image inpainting based on learning. In: Proceedings of The International MultiConference of Engineers and Computer Scientists, pp. 1442–1445 (2010)
13. Ma, J., Gu, X., Feng, C., Guo, J.: Study of thin cloud removal method for CBERS-02 image. Science in China Series E 48(2)(2005-03), 72–90 (2005)
14. Maalouf, A., Carre, P., Augereau, B., Fernandez Maloigne, C.: A bandelet-based Inpainting technique for clouds removal from remotely sensed images. IEEE Transactions On Geoscience And Remote Sensing 47(7), 2363–2371 (2009)
15. Salomon, D., Motta, G.: Handbook Of Data Compression. Springer, Heidelberg (2009)
16. Seow, M., Asari, V.: Ratio rule and homomorphic filter for enhancement of digital colour image. In: Proceedings of Neurocomputing, pp. 954–958 (2006)
17. Tasdizen, T., Whitaker, R., Burchard, P., Osher, S.: Geometric surface processing via normal maps. In: Proceedings of ACM Trans. Graph., pp. 1012–1033 (2003)
18. Zhang, X., Qin, F., Qin, Y.: Study on the thick cloud removal method based on multi-temporal remote sensing images. In: International Conference on Multimedia Technology (ICMT), pp. 1–3 (2010)

Hybrid Filter Based Simultaneous Localization and Mapping for a Mobile Robot

Amir Panah[1,2] and Karim Faez[3]

[1] Mechatronics Research Laboratory, Qazvin Islamic Azad University, Qazvin, Iran
[2] Young Researchers Club, Qazvin Islamic Azad University, Qazvin, Iran
[3] Electrical Engineering Department, Amirkabir University of Technology, Tehran, Iran
amir.panah@qiau.ac.ir, kfaez@aut.ac.ir

Abstract. A mobile robot autonomously explores the environment by interpreting the scene, building an appropriate map, and localizing itself relative to this map. This paper presents a Hybrid filter based Simultaneous Localization and Mapping (SLAM) approach for a mobile robot to compensate for the Unscented Kalman Filter (UKF) based SLAM errors inherently caused by its linearization process. The proposed Hybrid filter consists of a Multi Layer Perceptron (MLP) for neural network and UKF which is a milestone for SLAM applications. The proposed approach, based on a Hybrid filter, has some advantages in handling a robotic system with nonlinear motions because of the learning property of the MLP neural network. The simulation results show the effectiveness of the proposed algorithm comparing with an UKF based SLAM.

Keywords: Hybrid filter SLAM, MLP, SLAM, UKF.

1 Introduction

Currently, SLAM which is a relatively new subfield of robotics, is one of the most widely researched major subfields of mobile robotics. In order to solve SLAM problems, statistical approaches, such as Bayesian Filters, have received widespread acceptance [7]. Some of the most popular approaches for SLAM include using a Kalman filter (KF), an extended Kalman filter (EKF) and an unscented Kalman filter (UKF) and a particle filter [8]. The UKF SLAM makes a Gaussian noise assumption for the robot motion and its observation. In addition, the amount of uncertainty in the UKF SLAM algorithm must be relatively small; otherwise, the linearization in the UKF tends to unbearable errors. The UKF uses the unscented transform to linearize the motion and measurement models [13]. MLP neural network, adaptive to environmental information flowing through during the process, can be combined with an UKF to compensate for some of the disadvantages of an UKF SLAM approach [4],[12].

Qi Song and Yuqing He [9] in order to overcome the drawback of the normal unscented Kalman filter a novel adaptive UKF is developed and applied to nonlinear joint estimation of both time-varying states and modeling errors for helicopter. The filter is composed of two parallel master-slave UKFs, while the master UKF estimates the states/parameters and the slave one estimates the diagonal elements of the noise

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 343–352, 2011.

covariance matrix for the master UKF. Such a mechanism improves the adaptive ability of the UKF and enlarges its application scope.

Zhi Jun Yu et al [14] a new neural network aided Unscented Kalman filter is presented for tracking maneuvering target in distributed acoustic sensor networks. In this approach that using an offline neural network to correct these errors, the nonlinear inferring process is done by the normal Unscented Kalman filter. This method doesn't need complex modeling for tracking maneuvering target and is very suitable for real-time implementation.

Choi et al [2] solved the SLAM problem with a neural network based on an extended Kalman filter. According to the research results, the EKF SLAM based on Neural Network, shows better performance than the EKF SLAM.

Ronghui Zhan and Jianwei Wan [10] presents a robust learning algorithm for an multilayered neural network based on UKF. Since it gives a more accurate estimate of the link weights, the convergence performance is improved. This algorithm is then extended further to develop a neural network aided UKF for nonlinear state estimation.

In this paper, we present a Hybrid approach using MLP neural network and UKF based SLAM problem for decreasing uncertainty in compare to SLAM using UKF. We also discuss the effectiveness of MLP algorithm to handle nonlinear properties of a mobile robot.

Some related algorithms on SLAM are described in section 2, and the Hybrid SLAM algorithm is presented in section 3. Section 4 shows the simulation results of the SLAM based on UKF, and Hybrid filter. Concluding remarks, discussion and further research are discussed in section 5.

2 Related Algorithms for SLAM

2.1 Multi Layer Perceptron Neural Network

Frequently, neural networks are used especially in modeling and simulation of nonlinear systems. Neural networks have two fundamental characteristics of learning based on presentation experimental data and structural parallel. Specially, MLP which was evolved from single layer perceptron with a parallel processing pattern, has been proposed in the early days. MLP is suitable turn out for nonlinear information. The MLP with hidden layers with one or more input and output nodes, is used a typical feedforward neural network model used as a universal approximator. The output signals are generated through the homogeneously nonlinear function after summing signal values for each of the input nodes. In this process, signals are multiplied with appropriate weights and added with some bias values [5],[15].

2.2 Unscented Kalman Filter

This filter is built based on transformation as unscented transformation. In the UKF, there is no need to calculate Jacobian matrix. Since, the processing noise in this system is accumulative; therefore the augmented state vector is used to implementation this approach. In this approach, the mean and covariance estimation are calculated with considering the second order of the Taylor series [6].

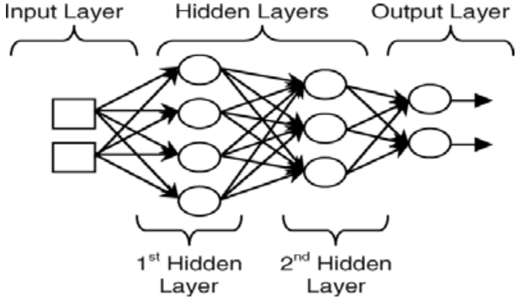


Fig. 1. A Multi Layer Perceptron Neural Network Structure

Suppose that a random variable x with mean μ and covariance P_x is given and also a random variable z related with x using: $z=f(x)$. Calculation Problem of the mean and covariance of z is the same as the predicted and corrected problem in UKF stages for a nonlinear system. In the Unscented Transformation method, a set of weighted points called sigma points are used to reach the mean and the covariance of random variable z. This sigma points should be selected in a way that have enjoy the mean μ and covariance P_x. For n-dimensional random variable with mean μ and covariance P_x, 2n+1 instance points are selected as follows. (0<i<n)

$$X_0 = \mu \quad , \quad W_0 = \frac{\lambda}{n+\lambda} \tag{1}$$

$$X_i = \mu + (\sqrt{(n+\lambda)P_x})_i \quad , \quad W_i = \frac{\lambda}{2(n+\lambda)} \tag{2}$$

$$X_{i+n} = \mu - (\sqrt{(n+\lambda)P_x})_i \quad , \quad W_{i+n} = \frac{\lambda}{2(n+\lambda)} \tag{3}$$

$$\lambda = \alpha^2(n+\beta) - n \tag{4}$$

N is the number of augment state. α and β are the coefficients that the estimation error can be minimized by adjusting them, and also their values influence on the error rate resulted from the higher terms in Taylor series. In the above mentioned equations, $k\in R$ and$(\sqrt{(n+\lambda)P_x})_i$, the i-th row or column of the matrix is the square root of $(n+\lambda)P_x$, W_i is the weight belongs to each point and k also is used for more accurate adjusting of UKF [6]. According to Unscented Transformation algorithm, each point in a set of points is first mapped to a new point by a nonlinear function, which results in a new set of sigma points. Then, we calculate the mean and the covariance values of the new random variable. Consider the following nonlinear system.

$$x_k = f(x_{k-1}, u_{k-1}, \varepsilon_k) \tag{5}$$

$$z_k = h(x_k, u_k, \delta_k) \tag{6}$$

Where x is the state vector and u is control input, ε is the system noise and δ is the measurement noise. In the first phase of implementing this filter, the augment state vector is formed as follows.

$$X_k^a = \begin{bmatrix} X_k \\ \varepsilon \\ \delta \end{bmatrix} \tag{7}$$

In the following, we will have all formulas used in the UKF which include two main sections, the Measurement Update and the Time Update [11].

- **The Time Update**

$$X_k^a = f(X_k^a, u_k, \varepsilon_k) \tag{8}$$

$$\mu_k = \sum_{i=0}^{2n} w_i X_{i,k}^a \tag{9}$$

$$P_k = \sum_{i=0}^{2n} w_i [X_{i,k}^a - \mu_k][X_{i,k}^a - \mu_k]^T \tag{10}$$

$$z_k = h(x_k, u_k, \delta_k) \tag{11}$$

$$\bar{z}_k = \sum_{i=0}^{2n} w_i z_k \tag{12}$$

- **The Measurement Update**

$$P_{\mathrm{x_k x_k}} = \sum_{i=0}^{2n} w_i [z_{i,k} - \bar{z}_k][z_{i,k} - \bar{z}_k]^T \tag{13}$$

$$P_{x_k y_k} = \sum_{i=0}^{2n} w_i [X_{i,k}^a - \mu_k][z_{i,k} - \bar{z}_k]^T \tag{14}$$

$$K_k = P_{x_k y_k} P_{x_k x_k}{}^{-1} \tag{15}$$

$$\mu_k = \mu_k + K_k(z_k - \bar{z}_k) \tag{16}$$

$$P_k = P_k - K_k P_{x_k x_k} K_k{}^T \tag{17}$$

Where X_k^a, μ_k, P_k, z_k, $\bar{z}_k$, $P_{x_k x_k}$, $P_{x_k y_k}$ and K_k, are defined as motion model, predicted mean, observation model, predicted observation, innovation covariance, cross correlation matrix and Kalman gain, respectively.

3 SLAM Algorithm Using Hybrid Filter

A new Hybrid filter SLAM using an UKF and a MLP is proposed here. The mean u_k which is derived from environmental information values $(xy\theta\varepsilon\delta)$ using the MLP algorithm, is entered to the prediction step, as shown in Fig. 2.

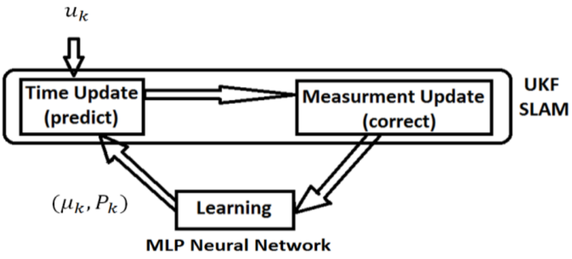


Fig. 2. The architecture of the Hybrid filter SLAM

In this paper, Basic inputs are mean, covariance which are calculated by prior input, u_{k-1}, and present input, u_k. The robot calculates the prior mean and covariance in The Time Update step and then in The Measurement Update step, calculates a Kalman gain, present mean and covariance and defined features.

3.1 Time Update (Predict)

In the following article, The Hybrid filter SLAM algorithm is described using a robot's pose and features, such as the location of landmarks. For the SLAM, the basic motion model of the mobile robot needs to be presented in the fallowing. A configuration of the robot with a state equation $X^a = (xy\theta\varepsilon\delta)^T$, has the form of Eq. (18).

$$X_k^a = \begin{bmatrix} x_k \\ y_k \\ \theta_k \\ \varepsilon_k \\ \delta_k \end{bmatrix} = \begin{bmatrix} x_{k-1} + v_k \Delta t cos(\theta_k) \\ y_{k-1} + v_k \Delta t sin(\theta_k) \\ \theta_{k-1+} v_k \Delta t sin(\frac{\Delta\theta}{L}) \\ \varepsilon_{k-1} \\ \delta_{k-1} \end{bmatrix} \tag{18}$$

$$u_k = v_k + N(0, M_k) \tag{19}$$

Where v_k is velocity of wheels, and L is the width between the robot's wheels, and Δt is the sampling period. Finally, M_k describes the covariance matrix of the noise in control space. The state equation for landmarks, is combined with the robot position, is denoted by the vector Y_k, where c is the number of landmarks. (0<i<c)

$$Y_k^a = \begin{bmatrix} X_k^a \\ m \end{bmatrix} = (\mathrm{x_k y_k \theta_k \varepsilon_k \delta_k} \quad \mathrm{m^i_{k,x} m^i_{k,y} s^i_k} 00)^\mathrm{T} \tag{20}$$

The state transition probability of a Hybrid filter SLAM has the form of Eq. (21).

$$X_k^a = f(X_{k-1}^a, u_{k-1}) + N(0, \varepsilon_k) \tag{21}$$

Under the linearity assumption where f represents the nonlinear functions, ε_k is the process noise, and u_k is control input. For the Taylor expansion of function, f its partial derivative is used with respect to X_k^a, as shown in Eq. (22).

$$f'(X_{k-1}^{a}, u_k) = \frac{\partial f(X_{k-1}^{a}, u_k)}{\partial X_k^{a}} \tag{22}$$

The continuing linearization using of f is approximated at u_k and u_{k-1} as shown in Eq. (23).

$$f(X_{k-1}^{a}, u_k) = f(\mu_{k-1}, u_k) + f'(\mu_{k-1}, u_k)(X_k^{a} - \mu_{k-1}) \tag{23}$$

With the replacement values obtained from equations 1, 2, 3, 4, 18, prior mean and covariance have the following form of:

$$\mu_k = \sum_{i=0}^{2n} w_i X_{i,k}^{a} \tag{24}$$

$$P_k = \sum_{i=0}^{2n} w_i [X_{i,k}^{a} - \mu_k][X_{i,k}^{a} - \mu_k]^T \tag{25}$$

$$z_k = \begin{bmatrix} \sqrt{(m_{k,x}^{i} - x_k)^2 + (m_{k,y}^{i} - y_k)^2} \\ tan^{-1}\left(\frac{m_{k,y}^{i} - y_k}{m_{k,x}^{i} - x_k}\right) - \theta_k \end{bmatrix} + N(0, \delta_k) \tag{26}$$

$$m^i = (m_x^i \quad m_y^i)^T \tag{27}$$

$$\bar{z}_k = \sum_{i=0}^{2n} w_i z_k \tag{28}$$

3.2 The Measurement Update (Correct)

To obtain the Kalman gain K_k, we need to calculate $P_{x_k x_k}$ and $P_{x_k y_k}$ in the feature based maps. To obtain the values $P_{x_k x_k}$ and $P_{x_k y_k}$, it is necessary to calculate X_k^a, μ_k, z_k, $\bar{z}_k$ that are calculated in equations 18, 24, 26, 28, with replacement of these values, we will have the following equations.

$$P_{x_k x_k} = \sum_{i=0}^{2n} w_i [z_{i,k} - \bar{z}_k][z_{i,k} - \bar{z}_k]^T \tag{29}$$

$$P_{x_k y_k} = \sum_{i=0}^{2n} w_i [X_{i,k}^{a} - \mu_k][z_{i,k} - \bar{z}_k]^T \tag{30}$$

$$K_k = P_{x_k y_k} {P_{x_k x_k}}^{-1} \tag{31}$$

In the following, complete combined MLP algorithm with UKF is described to SLAM of the mobile robot. MLP are involved with train through input data and measurement values. In the training process, weights are decided based on the relation of input data and each hidden layers. MLP Neural Network needs higher weight to objective value on the higher relations between poses and heading angle with comparing to measurement.

To apply a MLP, the mean values for each element are divided, and substituted by inputs of the MLP algorithm for each mean value. This research utilizes the MLP with two hidden layers, so the process equation is derived as Eq. (32). Under the assumption that this process does not have any bias, the n, n_1, n_2 and n_3 describe the number of input nodes, the first hidden layer's nodes, the second hidden layer's nodes and output layer's nodes with A, B and C, the number of nodes, respectively [8].

$$\begin{aligned}\hat{\mu}_k^n &= \xi\left[\sum_{n_3=0}^{C-1} w_k^{n_2n_3}\varphi_k^{n_2}\right] = \xi\left[\sum_{\gamma=0}^{C-1} w_k^{n_2\gamma}\xi\left[\sum_{B=0}^{B-1} w_k^{n_1n_2}\varphi_k^{n_1}\right]\right] \\ &= \xi\left[\sum_{n_3=0}^{C-1} w_k^{n_2n_3}\xi\left[\sum_{n_2=0}^{B-1} w_k^{n_1n_2}\xi\left[\sum_{n_1=0}^{A-1} w_k^{nn_1}\bar{\mu}_k^n\right]\right]\right] \qquad (32) \\ &(0 \le n_1 \le A-1, 0 \le n_2 \le B-1, 0 \le n_3 \le C-1)\end{aligned}$$

The next process to obtain the prior mean and the covariance is to update the results from Eq. (32). The process described in the above 5 steps repeats until the end of the navigation.

$$\mu_k = \hat{\mu}_k + K_k(z_k - \bar{z}_k) \qquad (33)$$

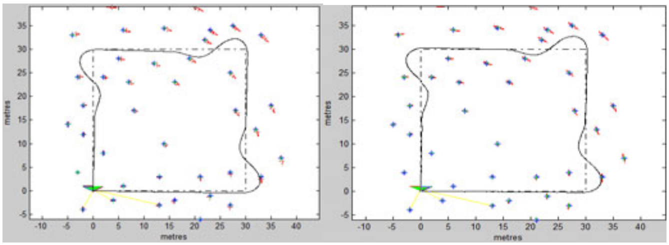
$$P_k = P_k - K_k P_{x_k x_k} {K_k}^T \qquad (34)$$

4 Simulations

To show the effectiveness of the proposed algorithm, the Matlab code, developed by Bailey [1], was modified. The simulation was performed with constraints on velocity, steering angle, system noise, observation noise, etc., for a robot with a wheel diameter of 1[m], maximum speeds of 3[m/sec], maximum steering angle and speed are 25[°] and 15[°/sec], respectively. The control input noise is assumed to be a zero mean Gaussian with σ_v(=0.2[m/s]) and σ_φ(=3[°]). For observation, the number of arbitrary features around waypoints was used. In the observation step, a range bearing sensor model and an observation model were used to measure the feature position and robot pose, which includes a noise with level of 0.1[m] in range and 1[°] in bearing. The sensor range is restricted to 15[m].In this research, a rectangular navigation case of the robot are surveyed. Specifications of navigation map are described in Table1.

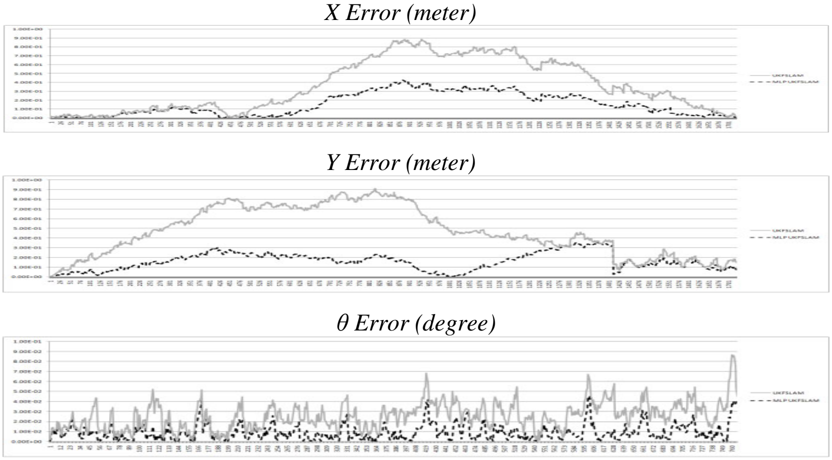
Table 1. Fundamental specification for navigation

Item	Rectangular
Feature	40
Waypoint	5
Area[m]	30*30



UKFSLAM MLP UKFSLAM

Fig. 3. Navigation result on rectangular map


X Error (meter)

Y Error (meter)

θ Error (degree)

Fig. 4. Navigation errors on rectangular map

4.1 Navigation on Rectangular Map

In the case of rectangular navigation, the UKF based on navigation and Hybrid filter based on navigation are shown in Fig. 3. The dashed line, show the paths of robots should traverse and the bold black line is Robot path, based on data described by the actual odometry. In Fig. 4, the gray bold line and the dashed black line are the x, y, and θ errors in the case of UKF SLAM and Hybrid filter SLAM, respectively.

5 Conclusions

The SLAM since the robot keeps track of its location by maintaining a map of the physical environment and an estimate of its position on that map, is one of the most fundamental problems in the quest for autonomous mobile robots. This paper proposes UKF SLAM based on MLP method for a mobile robot, to make up for the UKF SLAM error inherently caused by its linearization process and noise assumption. The proposed algorithm consists of two steps: the MLP Neural Network and the UKF algorithm. The simulation results show that the efficiency of the proposed algorithm based on MLP as compared with the UKF SLAM. To verify the effectiveness of the proposed algorithm, simulation in Matlab shown UKF SLAM has more errors than Hybrid filter SLAM. In addition, the simulation results confirm the Hybrid filter SLAM is more stable for robot navigation. In the future Research up on the robustness of the proposed algorithm, will verify under harsh and real-time condition using of fuzzy logic or structure change of neural network.

References

1. Bailey, T.: `http://www.personal.acfr.usyd.edu.au/tbailey`
2. Choi, M.Y., Sakthivel, R., Chung, W.K.: Neural network aided extended Kalman filter for SLAM problem. In: IEEE International Conference on Robotics and Automation, pp. 1686–1690 (2007)
3. Cho, S.H.: Trajectory Tracking Control of a Pneumatic X-Y Tabel using Neural Network Based PID Control. Int. J. Precis. Eng. Manuf. 10(5), 37–44 (2009)
4. Harb, M., Abielmona, R., Naji, K., Petriul, E.: Neural networks for environmental recognition and navigation of a mobile robot. In: IEEE International Instrumentation and Measurement Technology Conference, pp. 1123–1128 (2008)
5. Hu, Y.H., Hwang, J.N.: Handbook of Neural Network Signal Processing, pp. 3.1–3.23. CRC Press, Boca Raton (2001)
6. Julier, S.J., Uhlmann, J.K.: A New Extension of Kalman Filter to Nonlinear Systems. In: Proceedings of AeroSense: The 11th Int. Symp. on Aerospace/Defence Sensing, Simulation and Contro. (1997)
7. Kim, J.M., Kim, Y.T., Kim, S.S.: An accurate localization for mobile robot using extended Kalman filter and sensor fusion. In: IEEE International Joint Conference on Neural Networks, pp. 2928–2933 (2008)
8. Choi, K.-S., Lee, S.-G.: Enhanced SLAM for a Mobile Robot using Extended Kalman Filter and Neural Networks. International Journal Of Precision Engineering And Manufacturing 11(2), 255–264 (2010)
9. Song, Q., He, Y.: Adaptive Unscented Kalman Filter for Estimation of Modelling Errors for Helicopter. In: 2009 IEEE International Conference on Robotics and Biomimetics, Guilin, China, December 19-23, 2009, pp. 2463–2467 (2009)
10. Zhan, R., Wan, J.: Neural Network-Aided Adaptive Unscented Kalman Filter for Nonlinear State Estimation. IEEE Signal Processing Letters 13(7), 445–448 (2006)
11. Page, F.S.: Multiple-Object sensor Management and optimization. PHD thesis, in the faculty of Engineering, Science and Mathematics School of Electronics and Computer science (June 2009)
12. Vafaeesefat, A.: Optimum Creep Feed Grinding Process Conditions for Rene 80 Supper Alloy Using Neural network. Int. J. Precis. Eng. Manuf. 10(3), 5–11 (2009)
13. Zhu, J., Zheng, N., Yuan, Z., Zhang, Q., Zhang, X.: Unscented SLAM with conditional iterations. In: 2009 IEEE Intelligent Vehicles Symposium, pp. 134–139 (2009)

14. Yu, Z.-J., Dong, S.-L., Wei, J.-M., Xing, T., Liu, H.-T.: Neural Network Aided Unscented Kalman Filter for Maneuvering Target Tracking in Distributed Acoustic Sensor Networks. In: International Conference on Computing: Theory and Applications, Kolkata, India, March 5-7 (2007)
15. Zu, L., Wang, H.K., Yue, F.: Artificial neural networks for mobile robot acquiring heading angle. In: Proceedings of the Third International Conference on Machine Learning and Cybernetics, pp. 26–29 (2004)

Mitotic HEp-2 Cells Recognition under Class Skew

Gennaro Percannella[1], Paolo Soda[2], and Mario Vento[1]

[1] Dipartimento di Ingegneria dell'Informazione ed Ingegneria Elettrica,
Università di Salerno, Italy
[2] Facoltà di Ingegneria, Università Campus Bio-Medico di Roma, Italy
{pergen,mvento}@unisa.it, p.soda@unicampus.it

Abstract. Indirect immunofluorescence (IIF) is the recommended method to diagnose the presence of antinuclear autoantibodies in patient serum. A main step of the diagnostic procedure requires to detect mitotic cells in the well under examination. However, such cells rarely occur in comparison to other cells and, hence, traditional recognition algorithms fail in this task since they cannot cope with large differences between the number of samples in each class, resulting in a low predictive accuracy over the minority class. In this paper we present a system for mitotic cells recognition based on multiobjective optimisation, which is able to handle their low a priori probability. It chooses between the output of a classifier trained on the original skewed distribution and the output of a classifier trained according to a learning method addressing the course of imbalanced data. This choice is driven by a parameter whose value maximises, on a validation set, two objective functions, i.e. the global accuracy and the accuracies for each class. The approach has been evaluated on an annotated dataset of mitotic cells and successfully compared to five learning methods applying four different classification paradigms.

1 Introduction

Systemic autoimmune rheumatic diseases are characterised by the presence of antinuclear autoantibodies (ANAs), whose detection by indirect immunofluorescence (IIF) on HEp-2 slides [1,2] is required to confirm the diagnosis. Diagnostic procedure consists of the following four steps: image acquisition, mitosis detection, fluorescence intensity classification and staining pattern recognition (figure 1). Such tasks are very challenging for medical doctor, affecting the reliability of IIF diagnosis. Indeed, since IIF is a subjective and semi-quantitative method, physicians could be conservative or liberal during image classification on the basis of their skills and background, giving rise to classification variability [3]. Another significant reason of uncertainty is the low contrast of borderline and negative samples. In order to guarantee the correctness of the test, producers add to the slides some mitotic cells[1], which give to medical doctors the confidence

[1] Mitosis is the process by which a eukaryotic cell separates the chromosomes in its cell nucleus into two identical sets in two nuclei.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 353–362, 2011.

with their decisions. Indeed, first, medical doctors verify the correctness of well preparation process by detecting at least one fluorescent mitotic cell. Second, mitotic cells provide information on image staining pattern since they match with certain kinds of IIF patterns, including all stainings of antigens with different distribution throughout the cell cycle, such as midbody, CENP-F, mitotic splindle, centriole/centrosome and NuMA staining [4].

Since the demand of autoimmune laboratory tests has recently increased and ANAs detection in routine practice is far from being standardised [3], recent interests have been directed towards the development of computer-aided-diagnosis (CAD) systems supporting IIF diagnostic procedure. Investigated topics covered the areas of image acquisition [5,6], image segmentation [7,8,9] and fluorescence intensity classification [10] as well as staining pattern recognition [4,9,11,12,13].

Although mitotic cells detection plays a crucial role for CAD development being the first step in IIF diagnostic procedure, we find only one work on this topic [14]. It presented a set of features for mitotic cells recognition, testing the approach on an artificially balanced dataset composed of 126 samples. However, such cells rarely occur in a well in comparison to other cells giving rise to a skewed a priori samples distribution between mitotic and non-mitotic classes. In the following we assume that such classes correspond to minority and majority classes, respectively. Traditional classification algorithms cannot be successfully employed to recognise mitotic cells, since they are biased towards the majority class, resulting in poor predictive accuracy over the minority one. This happens because they are designed to minimize errors over training samples, ignoring classes composed of few instances [15,16].

Since the development of CAD system in IIF cannot prescind from the recognition of mitotic cells, and considering the skewed nature of the recognition problem, in the following we present a CAD system for mitotic cells recognition based on a method suited for learning under class skew, evaluating its performance on an annotated dataset of HEp-2 mitotic cells and achieving promising results. Our work differs from [14] since we consider here a strong skewed dataset reflecting the a priori distribution occurring in daily practice. We would like to address this lack found in the literature, opening the chance to develop fully automatic systems for IIF image analysis.

The paper is organised as follows: next session introduces classification approaches handling skewed distribution, Section 3 describes the dataset of HEp-2 mitotic cells and concisely recalls the set of descriptors used to represent the samples, Section 4 presents the experimental results and the discussion. Finally, Section 5 provides concluding remarks.

2 Classification

In this section, we first summarise recent researches on learning with class skew, and then present the approach we use.

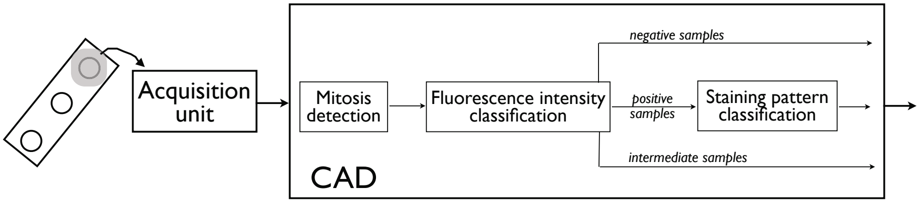


Fig. 1. Scheme of the diagnostic procedure for HEp-2 images

2.1 Classification Methods for Skewed Datasets

When the a priori samples distribution is skewed, traditional classification algorithms fail because they are designed to minimize errors over training samples, resulting in poor predictive accuracy over the minority one [15,16]. Existing learning approaches in these situations can be classified into the following five categories:

1. Under-sampling the majority class by resizing the training samples (TS), makes the class distribution more balanced [15,17]. This is done by sampling a subset N' from the majority training set N to make $|N'| = |P|$ where P is the minority class training set. The main drawback is the removal of potentially useful samples. One-sided selection is an under-sampling method that tries to overcome the above limitation removing borderline and redundant majority class samples, without touching minority class samples.
2. Over-sampling the minority class so as to match the size of the majority one, generates a set P' with $|P'| = |N|$ where P' is a combination of positive samples selected by this method, plus all positive samples in P. The main drawback of this approach is that it may increase the likelihood of over-fitting [15,17]. In order to overcome this problem, synthetic minority over-sampling technique was proposed [15]. It randomly creates synthetic samples in the feature space along the line segments to join any/all of the k minority class nearest neighbors [15].
3. Internally biasing the discrimination-based process to compensate class imbalance without altering the class distributions [16].
4. Multi-experts systems (MES), where each composing classifier C_i is trained on a subset of the majority class and on the whole minority class. After sampling several subsets N_i from N, C_i is trained on $N_i \cup P$. Then, outputs of C_i on test sample are combined to decide the final predications [18,19]. Indeed, a MES generally produces better results than those provided by its composing classifiers, avoiding drawbacks of both under and oversampling since each C_i is now trained on balanced sub-problems containing information on different aspects of N. Furthermore, in the framework of MES, in [18] the authors proposed the BalanceCascade method that supervisely explores the majority class using a cascade of base classifiers sequentially trained. As the aforementioned learning methods, MES approach produces better results on the minority class while harms the recognition of majority class samples.

5. Multiobjective optimisation: this method chooses between the output of a classifier trained on the original skewed distribution and the output of a classifier trained according to a learning method addressing the course of imbalanced data. This choice is driven by a parameter whose value maximises, on a validation set, two objective functions, i.e. the global accuracy and the accuracies for each class [20]. This approach would balance the recognition accuracies for each class, harming as less as possible the global accuracy. Indeed, other learning methods for class imbalance increase the accuracy over the minority class and decrease the accuracy over the majority one.

Finally, observe that the global recognition accuracy (acc) is sensitive to class skew and, hence, cannot be used as the only metric for performance estimation. In such cases, the geometric mean of accuracies ($gacc$) is also used to measure the classification performance [17]. It is defined as $gacc = \sqrt{acc^+ \cdot acc^-}$, where acc^+ and acc^- denote true positive and true negative rates, respectively. $gacc$ is a non-linear measure since a change in one of the two accuracies will influence the value of $gacc$.

2.2 Recognition Approach

Preliminary, we experimentally observe that learning algorithms numbered above as 1, 2, 3 and 4 provide unsatisfactory performance since they fail on most of the minority class samples, as shown in Table 1 and discussed in section 4. In order to overcome such an issue, we utilize a multi-objective optimization technique selecting the final output of the classification system between the output of a classifier trained according to a learning method addressing the course of class imbalance, and the output of a classifier adopting a training method for non-skewed data. This choice is driven by a parameter, referred to as threshold $\overline{t}$ in the following, whose value maximizes two objective functions, i.e. the global accuracy and the geometric mean accuracy.

The framework of our method is based on two classifiers, one trained on the original skewed distribution and another one built according to a class imbalance learning method. Given a test sample x, its label is given by:

$$O(x) = \begin{cases} O_{nSC}(x) & \text{if } \phi(x) \geq \overline{t} \\ O_{SC}(x) & \text{otherwise} \end{cases} \tag{1}$$

where $O_{nSC}(x)$ is the label assigned to sample x by any classifier trained on the original skewed distribution that does not apply any learning methods for skewed TS, which is referred to as nSC in the following. $O_{SC}(x)$ is the label assigned to sample x by any classifier trained according to a traditional learning method addressing the course of imbalanced TS, e.g. oversampling, undersampling, etc., which is named as SC in the rest of the paper. $\phi(x)$ is the reliability assigned by nSC to sample x and $\overline{t}$ is a real number in $[0, 1]$. The classification rule reported in 1 assigns the final label of x to the label returned by nSC when its reliability is larger than the threshold $\overline{t}$ because, in this case, it is reasonable to assume that nSC is likely to provide a correct classification. When $\phi(x)$ is below $\overline{t}$, $O(x)$

is equal to the label assigned by a classification system trained according to a method specifically tailored for imbalanced TS. Indeed, in this case the value of the reliability suggests that the decision returned by nSC should be not safe.

The value of the threshold $\bar{t}$ is set so that it maximizes both *acc* and *gacc* on a validation set, providing the best global performance as well as the most balanced accuracies on set disjoint from both training and test sets. The learning algorithm works as follows:

- Divide the labeled HEp-2 dataset $\mathbf{D}$ into training, validation and test sets, denoted by $\mathbf{D}_{Tr}$, $\mathbf{D}_{Va}$, $\mathbf{D}_{Te}$.
- Using $\mathbf{D}_{Tr}$, train both nSC and SC.
- Classify instances of $\mathbf{D}_{Va}$ with such classifiers.
- Apply equation 1 and measure both *gacc* and *acc* for each value of a threshold t ranging in $[0, 1]$. Indeed, *gacc* measures how much the accuracies over two classes are balanced, whereas *acc* estimates the global performance of the classification system.
- Build a graph where *gacc* and *acc* on the X and Y axis, respectively. The variation of t generates a set of points that can be used to plot a curve. Notice that curve extrema at $t = 0$ and $t = 1$ correspond to nSC and SC performance, respectively.
- The value $\bar{t}$ is given by $\bar{t} = \arg\min_t(||\mathbf{p}(t) - \mathbf{C}||)$, where $\mathbf{p}(t)$ is the pair of $gacc(t)$ and $acc(t)$ values measured on the validation set when the threshold t is used. $\mathbf{C} = (1, 1)$ is the ideal point in this plot since, the nearer the curve to this point, the better the performance obtained.

Notice that it is possible to proof that $\bar{t}$ is also an optimum solution of our problem according to multi-objective optimization theory [20].

3 Data Set and Features Extraction

Since, to our knowledge, there are not reference databases of IIF images publicly available, we populated a database of annotated mitotic cells with a resolution of 1388x1038 pixels, and a colour depth of 24 bits. The database consists of 28 positive images, showing the main six staining patterns shown in Fig. 3 [4]. Specialists manually segmented and annotated each cell from such images, reporting data on fluorescence intensity, pattern and mitosis phase. After this process, we get a mitotic dataset composed of 1527 cells, 70 mitotic cells and 1457 non mitotic cells. Non mitotic cells exhibits one of the six main staining patterns. The a priori probability of minority class is 4.8% and, hence, the dataset has a strong degree of imbalance in samples distribution.

Mitotic cells show peculiarities that can be exploited to get a set of features specifically tailored for this application. Such cells may exhibit two fluorescent patterns. The first, named as *negative mitosis*, has a fluorescent cell body, while the collapsed chromosomes mass located in the middle part of the cell does not exhibit a fluorescent pattern, or it has a weak fluorescence (panels A and B of Fig. 2). In the second pattern, referred to as *positive mitosis*, we observe the opposite situation, i.e. the cell body is weakly or not fluorescent, while the chromosomes mass is fluoroscent (panels C and D of Fig. 2). In both cases the collapsed chromosomes mass has a circular or elliptic shape, approximately. Comparing Fig. 2 and 3, we notice that such mitotic patterns are different from those of non mitotic cells.

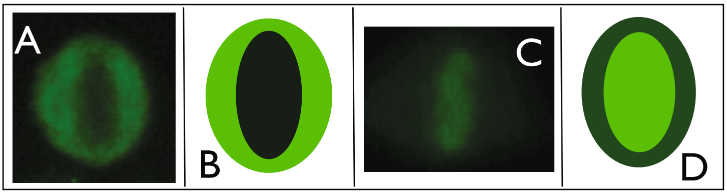


Fig. 2. Examples and stylised representations of positive (panels A-B) and negative (panels C-D) mitosis. Light and dark green represent high and low fluorescence.

In order to catch such peculiarities, we compute an an heterogeneous set of features belonging to morphological descriptors, texture measures, and local binary pattern (LBPs) [14].

Morphological descriptors look for elliptic shape and analyse the fluorescence intensity inside the cells. Information on the elliptic shape of chromosome mass, being either a positive or negative mitosis, is computed fitting a bivariate Gaussian distribution inside the cell body and deriving features based on the analysis of consecutive fittings. The second set of features consists of texture measures related to statistical and spectral descriptors. The third set of features is based on LBPs, which assign to each pixel of the image a label obtained comparing it with its neighbourhood matrix. The interested reader can refer to [14] for further details on used features.

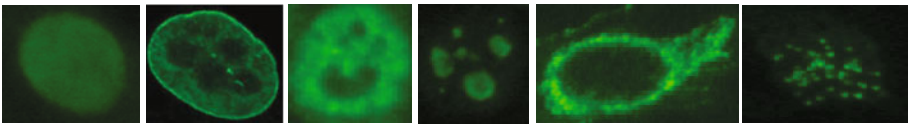


Fig. 3. Examples of the homogenous, peripheral nuclear, speckled, nucleolar, cytoplasmic and centromere staining patterns (left to right).

4 Results

The classification approach chooses the final output between the predictions of nSC and SC. As reported above, nSC is a classifier trained on the original skewed distribution that does not apply any learning method for imbalanced TS, whereas SC is a classifier trained according to a traditional learning method addressing the course of skewed TS, e.g. oversampling, undersampling, etc.

In order to analyse the classification performance, we test the following nine different classification configurations:

1. nSC: a classifier trained on the original skewed distribution which does not apply any learning methods for skewed TS;
2. SMOTE: a classifier trained according to synthetic minority over-sampling technique. In its implementation, we considered the three nearest neighbours, as suggested in [15];
3. OSS: a classifier trained according to one-sided selection [17];
4. MES-DCS: a multi-experts system where each classifier is trained on a subset $N_i \cup P$. According to results reported in [19], we apply random selection to sample N and dynamic classifier selection with local accuracy [21] to combine the outputs of base classifiers;
5. BaCa: a multi-experts system trained according to the BalanceCascade serial scheme [18];
6. nSC +SMOTE: a multi-objective optimisation technique choosing the final output between the predictions of nSC and SMOTE;
7. nSC+OSS: a multi-objective optimisation technique choosing the final output between the predictions of nSC and OSS;
8. nSC+MES-DCS: a multi-objective optimisation technique choosing the final output between the predictions of nSC and MES-DCS;
9. nSC+BaCa: a multi-objective optimisation technique RbB method choosing the final output between the predictions of nSC and BaCa.

Testing different configurations allows us to investigate different aspects of our recognition task. Indeed, configuration 1 permits us to estimate the performance achievable using a classifier trained on the original skewed distribution. Configurations 2-5 permit us to measure the performance achievable using "traditional" learning methods for skewed TS, whereas classifiers 6-9 are different applications of the proposed method. On this basis, we can compare the results of different classification methods on a strong skewed dataset.

We test four popular classifiers belonging to different paradigms. They are a Multi-Layer Perceptron (MLP) as a neural network, a k-Nearest Neighbour as a statistical classifier (kNN), a Support Vector Machine (SVM) as a kernel machine, and AdaBoost as an ensemble of classifiers. We report the results achieved using the best configuration of each classifier, set by preliminary tests.

Recognition performance, evaluated according to a 3-fold cross validation averaging out the final values of *gacc* and *acc*, are reported in Table1. The first part of the table reports the mean values of global accuracy (*acc*), which measures the global recognition rate. The second part of the table shows the mean values of geometric mean of accuracies (*gacc*), reflecting how much classifier performance are balanced. The third part of the table provides a synthtic comparison between the performance of nine classification schemes, considering together two previous performance metrics. To this aim, we compute for each classifier the L^2 distance between the average performance of each classification configuration and the ideal point $\mathbf{C}$ with coordinates $(1, 1)$ (see section 2.2). In order to facilitate the comparison between such distances, these values are normalised with

Table 1. Performance of tested classification configurations

Classifier	*Classification Performance (acc%)*								
	nSC	SMOTE	OSS	MES-DCS	BaCa	nSC+SMOTE	nSC+OSS	nSC+MES-DCS	nSC+BaCa
SVM	95.4	95.4	95.4	52.4	92.2	95.4	95.4	65.4	94.0
AdaBoost	95.4	94.8	94.2	57.4	94.6	95.1	94.7	74.8	95.3
MLP	95.4	94.5	93.3	55.3	93.1	94.5	94.9	74.7	94.4
3NN	95.2	92.9	84.0	54.6	92.4	92.7	84.5	74.6	92.5
Classification Performance (gacc%)									
SVM	0.0	0.0	0.0	52.8	23.5	0.0	0.0	52.9	25.4
AdaBoost	0.0	14.2	6.7	52.1	21.3	21.3	7.1	54.7	22.0
MLP	0.0	9.5	6.5	55.3	4.3	20.1	9.0	54.4	20.7
3NN	7.4	23.5	27.0	50.2	20.3	27.0	30.8	54.3	27.8
Normalised distance with ideal point **C** *()*									
SVM	70.8	70.8	70.8	47.4	54.4	70.8	70.8	41.3	52.9
AdaBoost	70.8	60.8	66.1	45.3	55.8	55.8	65.8	36.7	55.2
MLP	70.8	64.1	66.3	44.7	67.8	56.6	64.4	36.9	56.2
3NN	65.6	1591.0	52.8	47.7	56.6	51.9	50.1	37.0	51.3

respect to $\sqrt{2}$, which is the maximum distance from point **C**. Hence, such values range in [0,1]: the smaller the value, the more balanced the performance of the classifier.

Reported performance can be analysed from two points of view. First, we can compare the different learning methods on a high skewed dataset. Second, we can analyse the results considering the biomedical application at hand.

With reference to a machine learning perspective, we observe that nSC returns a value of *acc* equal to the a priori probability of majority class, thus misclassifying all minority class samples. This observation is confirmed by the value of *gacc*, which is equal to zero. This happens for all classifiers, other than 3NN. Introducing traditional learning methods for class imbalance, i.e. SMOTE, OSS, MES-DCS and BaCa, the situation improves since, now, *gacc* is different from zero. However, while *acc* keeps large, the value of *gacc* is still small, showing that such learning methods suffer with a so strong degree of imbalance. This does not happen using a SVM classifier in conjunction with learning methods for skewed data. Reason should lie in observing that such learning methods do not improve the linear separability of the samples induced by the chosen kernel in the features space. Let us now focus on the four applications of the proposed methods, i.e. nSC+SMOTE, nSC+OSS, nSC+MES-DCS and nSC+BaCa. With respect to the corresponding composing classifier, we observe that values of *acc* are approximately the same except for nSC+MES-DCS, whereas values of *gacc* increase. In case of nSC+MES-DCS, improvement of *gacc* counterbalances the decrease of *acc*. Turning our attention to the normalised distance from the ideal point, representing a perfect classification of both positive and negative samples, we notice that nSC+MES-DCS provides the best balanced performance, i.e. the best balance between the recognition rates of positive and negative samples. This happens for all tested classifiers, suggesting that such a scheme is the one best suited for this application. This observation introduces us to look at the results from a biomedical point of view. Results achieved by any nSC are clearly useless in a real scenario, since they fail on all positive samples which are the mitotic cells we would like to detect. The introduction of learning methods for skewed data improves a little the situation in comparison with nSC, since now *gacc* is larger than zero. However, in several cases, the values of *gacc* are still small, suggesting that such methods misclassify several positive samples. In contrast, the four applications of the proposed learning rule for the four different classifiers increase the value of *gacc*, implying that more mitotic cells are now correctly classified.

5 Conclusions

In this paper we have presented a classification approach for mitotic cells since their detection is a fundamental issue in developing a comprehensive CAD system in IIF. We have taken into consideration their low a priori probability, applying a learning rule which provides more balanced recognition rate with respect both to traditional classification methods as well as to other learning methods for class imbalance.

Future works are directed towards the integration of the system for mitotic cells recognition with systems for IIF image acquisition, fluorescence intensity classification and staining pattern recognition, according to figure 1.

Acknowledgments. This work has been supported by "Regione Campania" within the project "Classification of Immunofluorescence Images for the Diagnosis of Autoimmune Diseases" and by "Regione Lazio" under the programme "Intese con i Parchi scientifici e tecnologici del Lazio per favorire l'innovazione e il trasferimento tecnologico alle PMI laziali". The authors thank DAS s.r.l. for acquisition equipment and A. Rigon and L. Onofri for IIF images acquisition and annotation.

References

1. Kavanaugh, A., Tomar, R., et al.: Guidelines for clinical use of the antinuclear antibody test and tests for specific autoantibodies to nuclear antigens. American College of Pathologists, Archives of Pathology and Laboratory Medicine 124(1), 71–81 (2000)
2. Rigon, A., Soda, P., et al.: Indirect immunofluorescence in autoimmune diseases: Assessment of digital images for diagnostic purpose. Cytometry B (Clinical Cytometry) 72, 472–477 (2007)
3. Bizzaro, N., Tozzoli, R., et al.: Variability between methods to determine ANA, anti-dsDNA and anti-ENA autoantibodies: a collaborative study with the biomedical industry. Journal of Immunological Methods 219, 99–107 (1998)
4. Sack, U., Knoechner, S., et al.: Computer-assisted classification of HEp-2 immunofluorescence patterns in autoimmune diagnostics. Autoimmunity Reviews 2, 298–304 (2003)
5. Hiemann, R., Hilger, N., et al.: Objective quality evaluation of fluorescence images to optimize automatic image acquisition. Cytometry Part A 69, 182–184 (2006)

6. Soda, P., Rigon, A., et al.: Automatic acquisition of immunofluorescence images: Algorithms and evaluation. In: Computer Based Medical Systems, pp. 386–390. IEEE Computer Society, Los Alamitos (2006)
7. Huang, Y.L., Jao, Y.L., et al.: Adaptive automatic segmentation of HEp-2 cells in indirect immunofluorescence images. In: IEEE Int. Conf. on Sensor Networks, Ubiquitous and Trustworthy Computing, pp. 418–422 (2008)
8. Huang, Y.L., Chung, C.W., et al.: Outline detection for the HEp-2 cells in indirect immunofluorescence images using watershed segmentation. In: IEEE Int. Conf. on Sensor Networks, Ubiquitous and Trustworthy Computing, pp. 423–427 (2008)
9. Perner, P., Perner, H., Muller, B.: Mining knowledge for HEp-2 cell image classification. Journal Artificial Intelligence in Medicine 26, 161–173 (2002)
10. Soda, P., Iannello, G., Vento, M.: A multiple experts system for classifying fluorescence intensity in antinuclear autoantibodies analysis. Pattern Analysis & Applications 12(3), 215–226 (2009)
11. Hiemann, R., Büttner, T., et al.: Challenges of automated screening and differentiation of non-organ specific autoantibodies on HEp-2 cells. Autoimmunity Reviews 9(1), 17–22 (2009)
12. Hiemann, R., Hilger, N., et al.: Automatic analysis of immunofluorescence patterns of HEp-2 cells. Annals of the New York Academy of Sciences 1109(1), 358–371 (2007)
13. Soda, P., Iannello, G.: Aggregation of classifiers for staining pattern recognition in antinuclear autoantibodies analysis. IEEE Transactions on Information Technology in Biomedicine 13(3), 322–329 (2009)
14. Foggia, P., Percannella, G., et al.: Early experiences in mitotic cells recognition on hep-2 slides. In: 23rd IEEE Int. Symp. on. Computer-Based Medical Systems, CBMS 2010, pp. 38–43 (2010)
15. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16(3), 321–357 (2002)
16. Barandela, R., Sanchez, J.S., Garca, V., Rangel, E.: Strategies for learning in class imbalance problems. Pattern Recognition 36(3), 849–851 (2003)
17. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: One-sided selection. In: Machine Learning-International Workshop Then Conference, pp. 179–186. Morgan Kaufmann Publishers, Inc., San Francisco (1997)
18. Liu, X.Y., Wu, J., Zhou, Z.H.: Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 39(2), 539–550 (2009)
19. Soda, P.: An experimental comparison of MES aggregation rules in case of imbalanced datasets. In: 22nd IEEE Int. Symp. on Computer-Based Medical Systems, 2009, pp. 1–6 (2009)
20. Soda, P.: A multi-objective optimisation approach for class-imbalance learning. Pattern Recognition (2010) (in press)
21. Woods, K., Kegelmeyer, W.P., Bowyer, K.: Combination of multiple classifiers using local accuracy estimates. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 405–410 (1997)

Error Compensation by Sensor Re-calibration in Fringe Projection Based Optical 3D Stereo Scanners

Christian Bräuer-Burchardt, Peter Kühmstedt, and Gunther Notni

Fraunhofer Institute Applied Optics and Precision Engineering, Jena, Germany
{christian.braeuer-burchardt,peter.kuehmstedt,
gunther.notni}@iof.fraunhofer.de

Abstract. A new methodology for the determination and correction of calibration errors in optical 3D scanners is introduced. The accuracy of optical 3D scanners is reduced when calibration parameters are no longer valid caused by e.g. changes of temperature or mechanic influences. Whereas complete new calibration of a system may be impossible or too expensive, a compensation of the parameter errors can lead to acceptable results. The new method is based on measurement of a simple grid pattern or arbitrary measuring objects. Results show that the method provides acceptable results concerning point correspondence and scaling error compensation. It takes low effort and is easy to handle.

Keywords: camera calibration, epipolar geometry, fringe projection.

1 Introduction

High precision measuring systems based on image data require high precision optical components. Measuring accuracy, however, depends additionally on the quality of the geometric description of the components. Geometry determination of a 3D scanning system is performed in the process of camera calibration. Its correctness is crucial for the quality of a photogrammetric system and essential for its measuring accuracy.

Contactless 3D-measuring systems are increasingly used for industrial, medical, archaeological, and other applications. The requirements according measurement accuracy are growing while the handling should become easier. Modern photogrammetric measurement systems based on active structured light projection achieve measurement accuracies of up to 1:100000 compared to the length extension of the measuring field [1]. However, such high accuracies can only be achieved, if geometry of the measuring system is stable over time between calibration and measurement. This is, unfortunately, only the case, if certain measurement conditions strictly hold. However, in practical use this cannot always be ensured (see e.g. [2] or [3]).

Previous work has been dealt with the stability of camera calibration, e.g. by Läbe and Förstner [4], Habib et al. [5], and Rieke-Zapp et al. [3]. Zhang [6] gives an extensive review of the uncertainty of the epipolar geometry. A detailed description of the sensitivity of the 3D reconstruction depending on erroneous calibration parameters is presented by Dang et al. [7].

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 363–373, 2011.

Our goal was to develop a method which compensates measurement errors of optical 3D sensors caused by thermic instability or mechanic influences by correction of calibration parameters. Two novel methodologies for calibration analysis and correction are introduced. The first corrects those parameters leading to point correspondence errors. The second one additionally compensates scaling and deformation errors.

2 Measuring Principles

2.1 Phasogrammetry

Phasogrammetry is the mathematical connection of photogrammetry and fringe projection. The classical approach of fringe projection is described e.g. by Schreiber and Notni [8], which has been extended depending on several applications (see [9, 10]). The principle should be briefly explained as follows. A fringe projection unit projects one or two perpendicular, well defined fringe sequences onto the object, which is observed by one ore more cameras. These sequences may consist of a binary code sequence as the Gray code (see [11, 12]) and a sequence of up to 16 sinusoidal fringe patterns. The so called rough phase value [8], and in combination with the Gray code the unique phase value [11] is obtained using a sequence of sinusoidal pattern. Unique phase values are used to realize point correspondences in order to obtain measurement values by triangulation (see [1]).

2.2 Stereo Vision and Epipolar Geometry

Using active stereo vision, images of the object are captured from two different perspectives. Pairs of image coordinates resulting from the same object point (the homologous points) have to be identified. The object can be reconstructed by triangulation [1] using these points. In the case of active stereo vision a single intensity pattern or a sequence of patterns is projected onto the object under measure. There are several techniques to identify the homologous points in both cameras (see e.g. [1]).

Epipolar geometry is a well-known principle which is often used in photogrammetry when stereo systems are present. See for example [1]. It is characterized by an arrangement of two cameras observing almost the same object scene. A measuring object point M defines together with the projection centres O_1 and O_2 of the cameras a plane E in the 3D space. The images of E are corresponding epipolar lines concerning M. When the image point m of M is selected in the image I_1 of camera C_1, the corresponding point m_2 in the image I_2 of camera C_2 must lie on the corresponding epipolar line. This restricts the search area in the task of finding corresponding points.

In the following we assume a system consisting of two cameras C_1 and C_2 and one projector in a fixed arrangement.

2.3 Camera Calibration

Camera calibration means the determination of the intrinsic and extrinsic parameters (including lens distortion parameters) of an optical system. It has been extensively described in the literature, e.g. in [1, 10, 13]. Different principles are performed to conduct camera calibration. The selection depends on the kind of the optical system, the exterior conditions, the effort to be pushed, and the desired measurement quality. In case of the calibration of photogrammetric stereo camera pairs, the intrinsic parameters (principal length, principal point, and distortion description) of both cameras should be determined as well as the relative orientation between the cameras.

The position of the camera in the 3D coordinate system is described by the position of the projection centre O = (X, Y, Z) and a rotation matrix R obtained from the three orientation angles ω, ϕ, and κ. Considering stereo camera systems, the relative orientation between the two cameras (see [1]) should be considered, because the absolute position of the stereo sensor is usually out of interest.

Lens distortion may be considerable and should be corrected by a distortion correction operator D. Distortion may be described by distortion functions or by a field of selected distortion vectors (distortion matrix). The determination of D may be performed within the calibration procedure or separately (see e.g. [14]).

3 Fringe Projection Based 3D Stereo Scanners

3.1 Situation

Let a fringe projection based 3D stereo scanner following the principles described in the previous section be given. It takes one sequence of images from each measuring position. This requires a preliminary calibration of the sensor. Two application modes are distinguished.

First, two sequences of Gray code and sinusoidal images rotated by 90°, respectively, are projected and observed leading to two phase images (see [5]). An algorithm using phase correlation is used to identify the corresponding points in the image of camera C_2 to the given points with integer coordinates in the image of camera C_1. This mode should be denoted by m_1 mode.

Second mode, denoted by m_1 uses the epipolar geometry. The corresponding point q in the image of camera C_2 for a given point $p = (x, y)$ in the image of camera C_1 is found by search on the epipolar line defined by p and the relative orientation between the stereo camera pair. This requires the projection of only one sequence of Gray code and sinusoidal images. This mode should be denoted by m_2 mode. In the following we use mode m_2 for image acquisition as it is much faster compared to m_1.

Assume that all calibration parameters including distortion description D_1 and D_2 have been obtained by a suitable calibration process. The other parameters are X_1, Y_1, Z_1, ω_1, ϕ_1, κ_1, c_1, u_1, v_1, X_2, Y_2, Z_2, ω_2, ϕ_2, κ_2, c_2, u_2, and v_2, where (X_i, Y_i, Z_i) are the projection centres, ω_i, ϕ_i, and κ_i are the orientation angles, resulting in the rotation matrices R_i, c_i are the principal distances, and (u_i, v_i) are the principal points of the two cameras (i = 1, 2).

3.2 Stability of the Measurement

The measuring accuracy of a calculated 3D point using triangulation depends on several aspects. In the ideal case, calibration data is error free and positions of the found corresponding points are located perfectly. Then triangulation provides a proper intersection of the two rays r_1 and r_2 and the position error of the reconstructed point *M'* is zero. However, the calculated position *M'* will be erroneous due to the position error of *q* in the image of camera C_2 and the calibration data error. The position error of *q* depends on phase noise and a location error of the epipolar line. Whereas phase noise is a random error with typically normal distribution which can be statistically estimated, location error depends on the quality of calibration including distortion correction and the triangulation angle.

Because calibration is performed preliminary and is not updated during measurement cycles, the calibration parameters have to be stable in order to achieve constant good measurement accuracy. However, changes of temperature and mechanic influences as e.g. vibrations or shocks may considerably disturb the geometry of the sensor [3]. In that case current calibration data becomes erroneous leading to errors in the 3D data outcome. Unfortunately, the amount of these errors is difficult to estimate, e.g. in the case of mechanic shocks as the deviation of the calibration parameter is usually unknown. Performing a complete calibration update of the sensor is usually impracticable as it is time consuming, needs a sophisticated operator, and can often not be performed in the working position of the sensor. However, if the error of the current measurement is considerable, but unknown, the measurement result becomes useless.

In the case of fringe projection based 3D sensors operating temperatures often are significantly higher than the temperature of the environment, which might lead to a significant error. This error mainly results in a position error of corresponding points used for triangulation, and a scaling or deformation error due to invalid calibration parameter set. In the following section we introduce an approach to compensate for such errors using a low effort single shot measurement method.

Hand held, mobile, and moved sensors are additionally exposed to mechanic influences as shocks or vibrations which may disturb the calibration data.

4 Approach of Error Compensation

As mentioned above, there are two major problems concerning a disturbed sensor calibration. First, disturbed calibration leading to erroneous measurement results has to be detected. Second, considerable disturbances have to be corrected for. Hence our approach consists of two parts: detection and correction. The method should be applicable in the measurement modus of the sensor. It has not to require considerable effort, but easy to handle and being automated. It should realize a correction of calibration data such, that measurement errors are sufficiently reduced.

The approach was performed following two strategies. First, compensation procedure should be extremely simple and robust while the effort must be minimal. This leads to the concept of epipolar line correction (ELC). Second, calibration error has to be compensated by most accurate correction of the calibration parameters.

4.1 Simulation and Correction Model

To simplify the development of an automated method which compensates for errors due to thermal drift of the calibration parameters with low effort, not all of the 18 calibration parameters are considered. Instead, the most contributing parameters are selected. Analysis of the parameter error influence has been performed for identification. Let us consider first the epipolar line position error $err_{pos}(p,q)_i$, which should be defined as the perpendicular distance of the correct corresponding point q_i to the epipolar line defined by the set of calibration parameters. The mean epipolar line error ΔE_{mn} and the rms epipolar line error ΔE_{rms} of the image pair are defined by

$$\Delta E_{mn} = \frac{1}{n}\sum_{i=1}^{n} err_{pos}(p,q)_i \quad \text{and} \quad \Delta E_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(err_{pos}(p,q)_i\right)^2}\ , \tag{1}$$

respectively, where n is the number of corresponding point pairs which should be well distributed over the images. Additionally, the maximal epipolar line error ΔE_{max} may be considered: $\Delta E_{max} = \max\{\ |\ err_{pos}(p,q)_i\ |\}$.

The epipolar line error according to (1) is similar to the measure defined by Zhang to compare fundamental matrices obtained by different calibration procedures [6].

The influence of the different calibration parameters to ΔE_{rms} was determined as follows. A set of parameters similar to those of a real sensor [15] was constructed and a plane measuring object was simulated. This simulation provides well defined results of reconstructed 3D points. Then, a manipulation of every single calibration parameter was performed in order to estimate its influence on the mean epipolar line error ΔE_{rms}.

The set of parameters can be reduced due to symmetric effects of the parameters (e.g. an error of ΔX_1=0.1mm should have the same effect as $\Delta X_2 = -0.1$ mm) leading to $X_2, Y_2, Z_2, \omega_2, \varphi_2, \kappa_2, c_2, u_2, v_2$ as the considered parameters. In the following the index '$_2$' will be omitted. The nine parameters were disturbed with a reasonable error.

Table 1 shows the result of the mean epipolar line error and fig. 1 illustrates the distribution of selected error vectors. It can be seen that some of the parameters have a low effect on ΔE, e.g. X, ϕ, and κ. For some other parameters, the effects are very similar to each other (Y, ω, and v) concerning ΔE. Hence some of the parameters may be omitted when only the epipolar line error is considered in order to simplify the algorithm of finding the correction values. The effect of the parameter errors on the 3D data can be calculated by triangulation. It is also described by Dang et al. [7].

Unfortunately, epipolar line error ΔE is only sensitive in direction perpendicular to the epipolar lines. If the current calibration is disturbed by e.g. ΔX or $\Delta\phi$, this will hardly be detected by only analysing ΔE. Hence scaling error $\Delta S(P,Q)$ is defined by

$$\Delta S(P,Q) = \frac{\overline{PQ_{meas}} - \overline{PQ_{corr}}}{\overline{PQ_{corr}}} \tag{2}$$

where $\overline{PQ_{corr}}$ is the correct distance between two 3D points P and Q and $\overline{PQ_{meas}}$ is the measured distance between P and Q.

Analysis of the influence on scaling error was performed analogously to the analysis of the epipolar line error. The scaling error as function of the position of the points in the image was provided by simulation. Four regions (for the selected points P and Q) for the analysis were defined: horizontal above A, horizontal below B, vertical left C, and vertical right D, corresponding to long distant length measurements close to the margins of the measurement field. Mean absolute scaling error ΔS_{mna} with sum of non-negative weights equal to one ($w_A + w_B + w_C + w_D = 1$) is defined by

$$\Delta S_{mna} = \frac{w_A}{n_A}\sum_{i=1}^{n_A}\left|\Delta S_{Ai}\right| + \frac{w_B}{n_B}\sum_{i=1}^{n_B}\left|\Delta S_B i\right| + \frac{w_C}{n_C}\sum_{i=1}^{n_C}\left|\Delta S_{Ci}\right| + \frac{w_D}{n_D}\sum_{i=1}^{n_D}\left|\Delta S_{Di}\right| \cdot \tag{3}$$

Table 1. Epipolar line error ΔE depending on calibration parameter errors (examples)

parameter	X	Y	Z	ϕ	ω	κ	c	u	v
Δparameter	0.1mm	0.1mm	0.1mm	0.05°	0.05°	0.05°	0.1mm	0.01 mm	0.01mm
ΔE_{rms} [pixel]	0,03	2.62	0.28	0.08	3.88	0.11	0.65	0.02	2.08
ΔE_{max} [pixel]	0,07	2.66	0.42	0.14	3.93	0.20	1.28	0.05	2.12

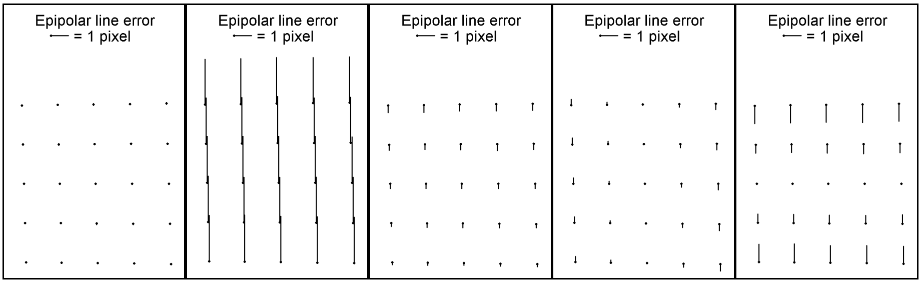


Fig. 1. Epipolar line error ΔE in dependence on selected single calibration parameter errors: ΔX_2= 0.1 mm, ΔY_2= 0.1 mm, ΔZ_2= 0.1 mm, $\Delta\kappa_2 = 0.1°$, and $\Delta c_2 = -0.1$ mm (from left to right)

4.2 Correction Approaches

For error compensation in a disturbed measurement we suggest two strategies. The first one should be denoted by epipolar line error correction (ELC) and can be obtained either by correction of the epipolar line positions or by calibration parameter manipulation. It shifts the epipolar lines according to the measured epipolar line error ΔE_{rms}. This leads to correct 2D coordinates of corresponding points resulting in prevention of locally appearing 3D errors. The advantage of this approach is a simple and fast determination, even applicable in the measurement process of arbitrary measuring objects. (The proposed grid pattern is not necessary!). However, scaling errors and deformations due to erroneous calibration cannot be corrected.

The second approach, denoted by calibration parameter correction (CPC), realizes a correction of all selected calibration parameters according to the minimization of ΔE_{rms} and ΔS_{mna} and includes the correction of point correspondences as well as the correction of scaling errors and deformations.

4.3 Selection of the Parameters

A set of relevant calibration parameters is selected according to the selected approach (ELC or CPC). This can be done after analysis of the errors ΔE and ΔS by performing a reference measurement. Alternatively, a heuristically found default parameter set (e.g. Y_2, Z_2, κ_2 or ω_2, Z_2, κ_2 for ELC and X_2, Y_2, Z_2, ϕ_2, ω_2 , κ_2 , c_2 for CPC) can be chosen.

The coordinates of the principal points u_1, v_1, u_2, and v_2 were omitted for both epipolar line error and scaling error estimation as the errors of u_1, v_1, u_2, and v_2 are assumed to be small in relation to those of the extrinsic parameters. Additionally, the remaining errors may be partly compensated by the selected correction parameters due to correlation effects.

4.4 Pure Determination of the Epipolar Line Error

In order to determine exclusively the epipolar line error, a measurement of an arbitrary measuring object is performed by the sensor using the m_1 mode. A number of points (e.g. 80 to 200) arranged in a rectangular grid covering the whole image field must be selected in the image of camera C_1 as reference point set. Selection criterion should be minimal position error probability. Erroneous point correspondences $(p, q)_i$ are determined on the epipolar lines. Second phase direction is used to estimate the correct point correspondences $(p', q')_i$. The epipolar line error is estimated by the difference of q_i and q'_i in direction perpendicular to the epipolar lines.

4.5 Determination of the Scaling Error

Scaling error determination is proposed as follows. A rectangular plane grid may be used with known exact distances between the grid points. The coordinates of the grid points are determined with high sub-pixel accuracy. Length measurements l_i are performed using selected points. Measurement results are compared to the correct distances l_i^{ref} leading to four values ΔS_A, ΔS_B, ΔS_C, and ΔS_D of ΔS according to (3).

4.6 Determination of the Correction Parameters

The set of point correspondences $(p, q)_i$ obtained by the reference measurement, and four point correspondences for length measurement are the input for the algorithm of the correction parameter estimation. A test quantity $\Delta T = f_E \cdot \Delta E_{rms} + f_S \cdot \Delta S_{mna}$ with $f_E + f_S = 1$ to be minimized is defined, where f_E and f_S are meaningful, heuristically set scaling factors which can be obtained by experiments. The accuracy of the scaling error determination depends on the precision of the used grid and the grid point localization accuracy. If scaling error ΔS_{mna} should be neglected, $f_S = 0$ may be set leading to reduced parameter set and faster re-calibration.

The selected m parameters are changed systematically by addition of $k \cdot \delta par_j$, j=1,…,m, where $k = -1, 0, +1$, and $\delta par_j = (\delta X, \delta Y, \delta Z, \delta\phi, \delta\omega, \delta\kappa; \textit{ and } \delta c)$, respectively. The initial value for the δpar_j are set to meaningful, heuristically obtained values, e.g. 0.1 mm for X, Y, and Z, and 0.01° for ω, ϕ, and κ. The test quantity ΔT is determined for all 3^m combinations of $k \cdot \delta par_j$. The minimum ΔT_{min} = min $\{\Delta T_j (p,q)_i\}$, j=1,…, 3^m defines the favourite combination of the manipulated parameters. Depending on whether it holds $k_j \neq 0$ or $k_j = 0$ the δpar_j do not change or are devided by two, respectively: $\delta par_j := \delta par_j / 2$. The calculation of ΔT is iterated until no improvement of ΔT occurs or the maximum number of iterations is achieved.

4.7 Selection of the Correction Strategy

According to the factors f_E and f_S the influence of the epipolar line error or the scaling error, respectively, dominates. Actually, the use of only the scaling error (i.e. $f_E = 0$) would be sufficient in order to correct the calibration parameters. However, $f_E = 0$ leads to worse results than in the case $f_E > 0$ which has been verified by experiments (see next section). Additionally, using fringe projection systems the epipolar line error can be usually determined with low uncertainty. The accuracy of ΔS determination depends on the precision of the grid pattern measurement and may be lower than determination of ΔE. Hence epipolar line error should be always considered.

5 Results

5.1 Analysis of Thermic Behaviour

First we analysed a 3D stereo sensor (called DS) based on fringe projection which is used for intraoral measurements [15]. It observes a measurement field of about 20 mm x 15 mm. The depth of the measuring volume is about 12 mm. Image size is 516 x 778 pixel. Due to the absence of active ventilation and cooling the electronic components lead to a significant increase of the operating temperature to more than 30°C. Hence an influence of the changing temperature on the measuring accuracy is expected.

The following measurements were performed. First, only the epipolar line error was investigated. Therefore a white plane surface was measured using mode m_1 (see section 3.1). Epipolar line error ΔE_{rms} was determined and the estimation of the correction parameters Y_2, Z_2, and κ_2 was performed as described in section 4. The resulting parameter values were used to correct the location of the epipolar lines. The results before and after correction for one selected but representative measurement are documented in fig. 2 (left).

The same procedure was performed using the reference grid object. Parameters X_2, ω_2, ϕ_2, and c_2 were additionally included into the compensation. Results are given in table 2 and illustrated in fig. 2 (right). Note that the temperature caused (uncorrected) errors ΔE and ΔS are minimal at about 30 min after switching on the system. Thereafter, uncorrected errors increase again. After correction, however, epipolar line error is always below 0.06 pixel, and scaling error is below 0.15 %.

Table 2. Thermic drift influenced ΔE and ΔS measurement. $\Delta E'$ and $\Delta S'$ are corrected values

time \ error	ΔE_{rms} [pixel]	$\Delta E'_{rms}$ [pixel]	ΔE_{max} [pixel]	$\Delta E'_{\max}$ [pixel]	ΔS_{mna} [%]	$\Delta S'_{mna}$ [%]
0 min	0,68	0,06	0,85	0,31	1,38	0,14
10 min	0,08	0,03	0,23	0,17	0,90	0,12
20 min	0,10	0,02	0,15	0,10	0,25	0,09
30 min	0,05	0,05	0,17	0,15	0,07	0,06
60 min	0,36	0,03	0,50	0,15	0,08	0,06
90 min	0,44	0,03	0,53	0,09	0,11	0,10

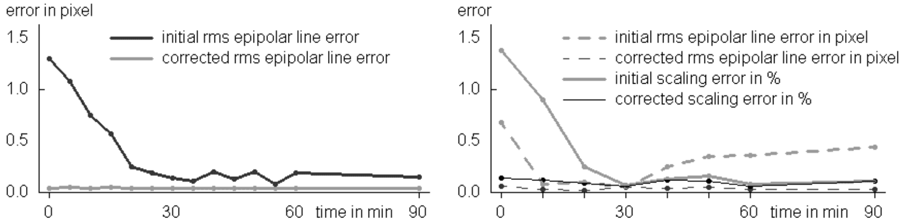


Fig. 2. Temperature influenced error ΔE (left) and both ΔE and ΔS (right) of DS over 90 min

5.2 Detection of Mechanic Influences

Next, the long-term behaviour of the stability of the calibration data of different flexible sensors was analysed. Evaluation of ΔE_{rms} was performed at different days over a period of several weeks. Exceptions after certain (but unknown) events are possible which require a sensor re-calibration as proposed. See an example in fig. 3 showing a relative high stability over three weeks. Then, a (probably) mechanic shock disturbed the calibration of the sensor at day 21. The effect of this change to a measurement of a prism in the unhandled 3D measurement and corrected result is also shown in fig. 3.

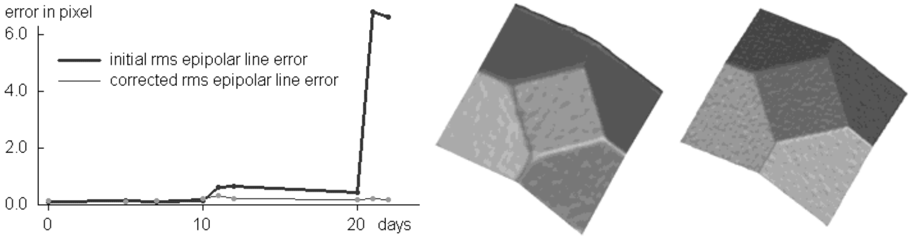


Fig. 3. Result of long-term behavior of ΔE (left) and influence of the sudden error on the 3D measurement of a prism: erroneous result with blurred edges (middle) and result after re-calibration (right)

5.3 Weight Optimization

As already mentioned in section 4.7, consideration of ΔS leads to ΔE being redundant. However, experiments with omission of ΔE did not show the desired results. Hence, experiments with varying ratio between f_E and f_S were performed using a representative dataset of our scanner DS. A three-parameter (Y_2, Z_2, κ_2) and a seven-parameter (X_2, Y_2, Z_2, ϕ_2, ω_2, κ_2, c_2) manipulation of the calibration parameters were performed.

The results are documented in table 3. Initial epipolar line error was $\Delta E_{rms} = 0.68$ and scaling error 1.39 according to (1) and (2). Whereas three-parameter manipulation gives almost constant ΔS error and a low ΔE value after correction for values $f_E > 0.1$ and considerable ΔE error for $f_E \le 0.1$. Seven-parameter manipulation gives acceptable results in the range of $0.1 \le f_E \le 0.9$. Hence, if ΔS determination is omitted, three-parameter manipulation should be performed and ΔE must not be omitted at all. If scaling error can be determined sufficiently accurate, seven-parameter manipulation should be performed with about $f_E = 0.3$ and $f_S = 0.7$.

Table. 3. Epipolar line and scaling error after three-parameter (Y_2, Z_2, κ_2) manipulation (left) and seven-parameter (X_2, Y_2, Z_2, ϕ_2, ω_2, κ_2, c_2) manipulation (right) depending on f_E and f_S

f_E	f_S	ΔE^3_{rms} [pixel]	ΔS^3_{mna} [%]	ΔE^7_{rms} [pixel]	ΔS^7_{mna} [%]
0.00	1.00	51,85	0,99	14,00	0,10
0.01	0.99	0,99	1,05	0,13	0,10
0.10	0.90	0,88	1,06	0,11	0,10
0.30	0.70	0,07	1,37	0,10	0,11
0.50	0.50	0,10	1,43	0,06	0,14
0.70	0.30	0,10	1,43	0,04	0,16
0.90	0.10	0,05	1,38	0,04	0,17
0.99	0.01	0,06	1,39	0,04	0,18
1.00	0.00	0,06	1,39	0,04	4,00

6 Summary, Discussion, and Outlook

A new methodology for sensor re-calibration and compensation calibration errors of stereo 3D scanners based on fringe projection technique was introduced. These errors are due to the thermic state of the sensor or mechanic influences. The method requires minimal effort and allows a sufficient correction of the distorted measuring results. The technique needs only a one position measurement of an arbitrary object or a grid pattern with known lengths between certain points, respectively.

The results show, that thermic drift or mechanic shocks may lead to considerable errors in the point correspondence finding which may be corrected adequately.

Due to high correlation between the calibration parameters, re-calibration based on the proposed method does not necessarily find the set of "true" parameters. Hence, further analysis of the parameter influence and searching for a better separation of these influences will be addressed in further work. Additionally, the proposed method should be tested on more different scanner systems.

References

1. Luhmann, T., Robson, S., Kyle, S., Harley, I.: Close range photogrammetry. Wiley Whittles Publishing, Chichester (2006)
2. Hastedt, H., Luhmann, T., Tecklenburg, W.: Image-variant interior orientation and sensor modelling of high-quality digital cameras. IAPRS 34(5), 27–32 (2002)
3. Rieke-Zapp, D., Tecklenburg, W., Peipe, J., Hastedt, H., Haig, C.: Evaluation of the geometric stability and the accuracy potential of digital cameras – Comparing mechanical stabilisation versus parameterisation. ISPRS Journal 64/3, 248–258 (2009)
4. Läbe, T., Förstner, W.: Geometric Stability of Low-Cost Digital Consumer Cameras. In: Proc. ISPRS, pp. 528–535 (2004)
5. Habib, A.F., Pullivelli, A.M., Morgan, M.F.: Quantitative measures for the evaluation of camera stability. Opt. Eng. 44, 033605-1– 033605-8 (2005)

6. Zhang, Z.: Determining the epipolar geometry and its uncertainty: a review. IJCV 27(2), 161–198 (1998)
7. Dang, T., Hoffmann, C., Stiller, C.: Continuous stereo self-calibration by camera parameter tracking. IEEE Transactions on Image Processing 18(7), 1536–1550 (2009)
8. Schreiber, W., Notni, G.: Theory and arrangements of self-calibrating whole-body three-dimensional measurement systems using fringe projection technique. Opt. Eng. 39, 159–169 (2000)
9. Reich, C., Ritter, R., Thesing, J.: 3-D shape measurement of complex objects by combining photogrammetry and fringe projection. Opt. Eng. 39, 224–231 (2000)
10. Chen, F., Brown, G.M.: Overview of three-dimensional shape measurement using optical methods. Opt. Eng. 39, 10–22 (2000)
11. Sansoni, G., Carocci, M., Rodella, R.: Three-dimensional vision based on a combination of Gray-code and phase-shift light projection: Analysis and compensation of the systematic errors. Applied Optics 38(31), 6565–6573 (1999)
12. Thesing, J.: New approaches for phase determination. In: Proc. SPIE, vol. 3478, pp. 133–141 (1998)
13. Brown, D.C.: Close-range camera calibration. Photogram. Eng. 37(8), 855–866 (1971)
14. Bräuer-Burchardt, C.: A simple new method for precise lens distortion correction of low cost camera systems. In: Rasmussen, C.E., Bülthoff, H.H., Schölkopf, B., Giese, M.A. (eds.) DAGM 2004. LNCS, vol. 3175, pp. 570–577. Springer, Heidelberg (2004)
15. Kühmstedt, P., Bräuer-Burchardt, C., Munkelt, C., Heinze, M., Palme, M., Schmidt, I., Hintersehr, J., Notni, G.: Intraoral 3D scanner. In: Proc. SPIE, vol. 6762, 67620E (2007)

Advanced Safety Sensor for Gate Automation

Luca Bombini, Alberto Broggi, and Stefano Debattisti

VisLab – Dipartimento di Ingegneria dell'Informazione
Università degli Studi di Parma
{broggi,bombini,deba}@vislab.it
http://www.vislab.it

Abstract. This paper presents a vision-based method able to increase safety in access automation systems. These systems include automatic swing and sliding gates, bollards and barriers to prevent unwanted access. In current installations the anti-crushing protection is ensured by an electronic device installed on the control boards, which directly controls drive torque, and a couple of infrared photodetectors: when an obstacle is touched by the gate leafs or barriers, or cuts the infrared beam, the control board stops the gate movement. Conversely, the new method proposed in this paper avoids collisions with obstacles and increases the overall detection performance. The new device returns a stop signal when an obstacle is present in a predefined area. The algorithm has been integrated in a real access system to test its performance.

Keywords: safety sensor, gate automation control, obstacle detection, stereo vision.

1 Introduction

Access automation systems are used in many settings including residential or public areas, like park systems. They include automatic gates and barriers, swing and sliding gates, bollards and barriers to prevent unwanted access or to regulate traffic flow. The safety of these installations is a primary requirement considering their very large distribution in residential and commercial areas. Some of the possible hazards associated with this kind of automation are shown below:

* Crushing during closing.
* Shearing.
* Impacts.
* Person lifting.
* Entrapment.

In response to the latest European standards concerning safety in the access automation systems[6], industries developed a lot of products and accessories that comply with the latest European directives.

Despite that, these products have some key limitations due to the technology used: obstacle recognition is usually done with photodetectors, performing

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 374–383, 2011.

obstacle localization only along the straight line that connect transmitter to receiver; for this reason this kind of sensors can not cover the whole danger area. Another key limitation of photodetectors is that doors themselves must not be detected as obstacles, so photodetectors must be placed outside the danger area.

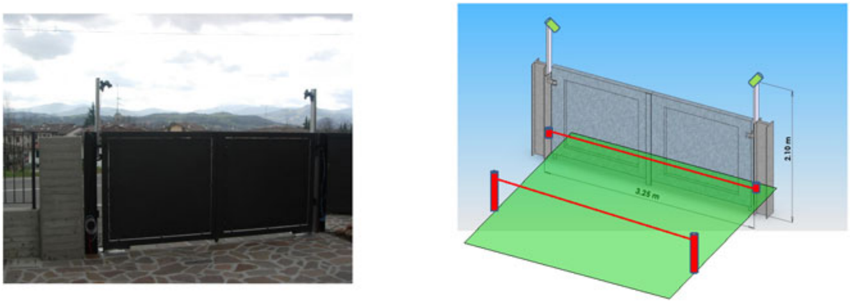
In order to increase the safety in access system, this paper shows a method able to meet the safety European specification and therefore avoid the hazards mentioned before. The goal of this work is to create a vision-based system able to detect obstacles in the danger area; in this way, the whole gate maneuver area can be monitored by the safety system.

The tests setup selected for the experimentation (see fig.1), include an automatic gate with two gate leafs. We have selected this scenario in order to deal with the most difficult case: two filled swing gate leafs that partially occluded the danger area in the opening/closing movement.

In order to obtain a robust system, several problems have to be analyzed and exploited:

- **Different light conditions:** the automatic gate access is placed outdoor and it has to work day and night; cameras can also be dazzled by the sun during the day.
- **Vibrations:** the mechanical movement of the doors can introduce vibrations in the gate pillars that would be transmitted to the cameras.
- **Filled Doors:** the danger area is partially occluded by the gate leafs.
- **Camera positioning:** the cameras are placed on the gate pillars (see Fig.1) in order to increase the field of view of the stereo systems. This positioning is unsual for stereo-based algorithms.

A very general setup has been studied to fit different kinds of automatic gates. A pair of CMOS cameras are mounted on the gate pillars overlooking the danger area. The cameras are connected to an embedded computer that processes the images and directly drives the access control board.



(a) *(a)*

Fig. 1. (a) The camera installed on an automatic gate. (b) The cameras (green) are placed on the pillars and can replace the photodetector (red). The area monitored by the new vision-based system (green plane) is larger than the one covered by the photodetectors (red line).

2 The Algorithm

The main goal of the application described in this paper is obstacle detection in a pre-defined area. Obstacle detection is a well known research field and several systems and solutions have been developed during the last years. The proposed approach differs from all the other existing systems because it takes in account the presence of a moving gate leafs in the analyzed scene. Therefore described algorithm will focus mainly on the innovative gate leafs detection system instead of the obstacle detection.
The gate leafs create two major problems:

- **Obstacle Detection:** gate leafs are not obstacles that need to stop the gate movement; we have to recognize them in order to avoid any possible obstacle false positive detection.
- **Occlusion:** we are working on a gate with filled swing-gate, so the gate leafs may occlude the cameras field of view.

In order to solve these problems, we have decided to create four virtual cameras, out of images coming from the two real cameras. The virtual cameras will be analyzed in the next subsection.

The whole algorithm is pictured in fig.2 and is performed in three main steps:

* Lens distortion and perspective removal from both images.
* Stereo Image Segmentation and creation of the images belonging to the four virtual cameras.
* Obstacle Detection.

Concerning the first step, during an offline preprocessing, a lookup table (LUT) that allows a fast pixel remapping is generated; this LUT associates each pixel in the distorted image to its homologous pixel on the undistorted image. Images of

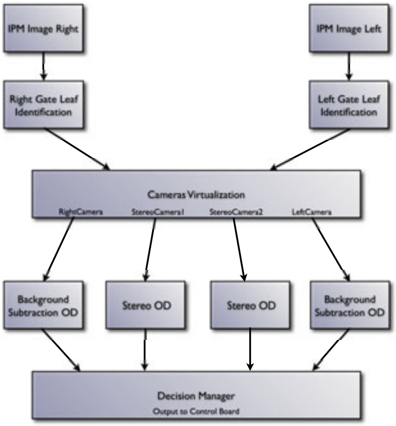


Fig. 2. Algorithm scheme

a grid, painted on the danger area, are used to compute the LUT and a manual system to pinpoint all the crossing points on the source image is used. Thanks to the knowledge of the relative position of the cameras with respect to the grid itself and to the assumption that the danger area can be considered nearly flat, it is possible to compute a new image (the IPM image) removing both the perspective effect and camera distortion at once. A nonlinear interpolation function is used to remap the pixels of the source image that are not crosspoints [1].

Gate leafs recognition is performed next, as described in section 2.2. Recognizing the exact angular position of both gate leafs allows image segmentation in different zones with different characteristics. In particular the images have been segmented in stereo and mono zones, representing image areas visible from both cameras or only one camera.

Once images have been segmented, the algorithm performs camera virtualization: for each image area that will be analyzed by the obstacle detection algorithm, a virtual camera is created and the obstacle detection is executed directly on the images coming from this virtual sensor. In particular, as we will see in section 2.3, different approaches on obstacle detection have been used for different virtual cameras, due to the different information provided by each virtual sensor.

Therefore, obstacle detection based on background subtraction is applied to the images coming from the mono virtual cameras, while an obstacle detection algorithm that exploits the difference between IPM images is applied to the images coming from the stereo virtual cameras.

Each obstacle detection algorithm provides an obstacles map of the analyzed area as output, with some other information like obstacle position and size. These maps are then merged by a decision manager (DM) in order to provide a unique command to the gate control. In particular the DM analyzes and compares information provided by available maps according to the status of the gate as explained in table 1. The virtual camera names of table 1 are explained in section 2.1.

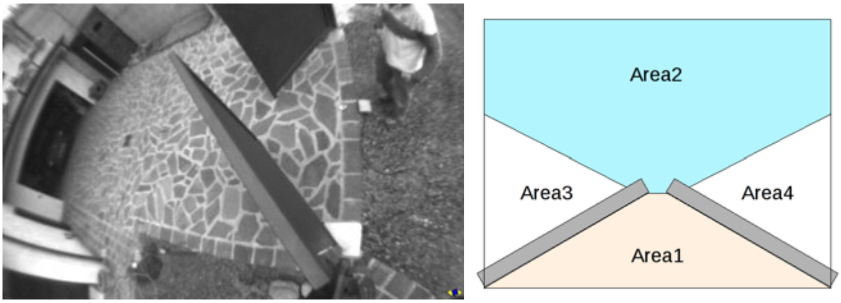
Table 1. Decision Manager

GateStatus	Active Virtual Sensors	Note
Opening	LC/RC, SC1	Objects placed behind the gate opening line are not considered as obstacles
Closing	SC1, SC2	Objects placed in front of the gate opening line are not considered as obstacles
Stopped	LC, RC, SC1, SC2	All the surveilled area is considered as dangerous

2.1 Virtual Sensors

As seen in fig.2, the second step of the algorithm is image segmentation and virtual cameras creation.

In fig.3.a we can frame a typical gate-moving situation: the gate leafs are opening and the two cameras can not entirely see the safety zone.



(a) *(b)*

Fig. 3. Image from camera Left (a) and danger area segmentation (b)

The basic idea of our image segmentation is to split the original images in several sub-images with different characteristic, as we can see in fig.3.b.

The areas pictured in fig.3.b require a different analysis: *Area1* and *Area2* are always visible from both cameras, while *Area3* and *Area4* are visible from only one camera during the gate opening movement. Therefore, the following 4 virtual cameras are created:

- **LeftCamera (LC):** includes the *Area3* image zone by segmenting the images coming from the left camera.
- **Right Camera (RC):** includes the *Area4* image zone by segmenting the images coming from the right camera.
- **Stereo Camera 1 (SC1):** includes the *Area1* image zone by segmenting the images coming from both cameras.
- **Stereo Camera 2 (SC2):** includes the *Area2* image zone by segmenting the images coming from both cameras.

Each virtual vision system provides images to a specific obstacle detection algorithm.

2.2 Gate Leafs detection

At the base of the Gate Leafs Detection System (GLDS) there is the knowledge of the gate leafs' measures: starting from these values a mathematical model of each gate leafs is created, typically as a regular parallelepiped. The projection of the mathematical model on the IPM images coming from the calibrated cameras reduces the gate leafs detection problem to the detection of the gate leafs' angular position. Angular position may be detected using an artificial vision algorithm or using an encoder integrated on leafs electric motors. The encoder-based solution is safer than the other one, and is generally more precise; on the other hand using an encoder require a dedicated hardware component. For this reason, an artificial vision algorithm has been developed, but, because of the great differences between different types of gates, this algorithm is suitable only for systems described in Sec. 1 . The main steps of $GLDS$ are shown below:

- **Background subtraction and binarization:**moving objects are extracted from the images.
- **Detection window reduction:** the size of the images on which the detection will be performed is reduced, mantaining only the interesting areas.
- **Comparation with gate leafs model:** the objects found in the images are compared with the gate leafs mathematical model in order to estimate their angular position.
- **Tuning of the detection:** the angular position is refined using some assumptions.

The first step of the algorithm creates binarized images highlighting the objects moving in the scene; to do this, a standard background subtraction algorithm [2] has been used. The images coming from this step are next filtered with a Sobel filter and then binarized. In order to reduce the detection window we made some assumptions: we assume the gate trajectory as continuos, although not perfectly predictable, and the gate leafs velocity as quasi-constant. With these assumptions, and knowing the gate leaf position in the last frame, we are able to estimate the current gate leaf position; then we can consequently reduce the detection window to a sub-image, closer to the position we predicted. The detection window reduction brings several advantages: first, algorithm processing time improves, since the most complex part of the algorithm is represented by the comparison of the detection window with the mathematical gate leafs model; furthermore the reduction of the detection window size avoids many possible false positive detections, because the main component of the resulting image is represented by the gate leafs' edges. The output of this step of the algorithm is represented by an image, called detection window, containing the edges of the objects moving in the scene; to find out the position of the gate leafs we compare this image with the mathematical model of the gate. To do this, all the white pixels contained in the parallelepiped that represents a gate leaf are accumulated; this value represents a performance index of the correctness of the estimated angular position. This operation is repeated within a range of about 10 degrees with a step of 1 degrees, as we can see in fig.4.

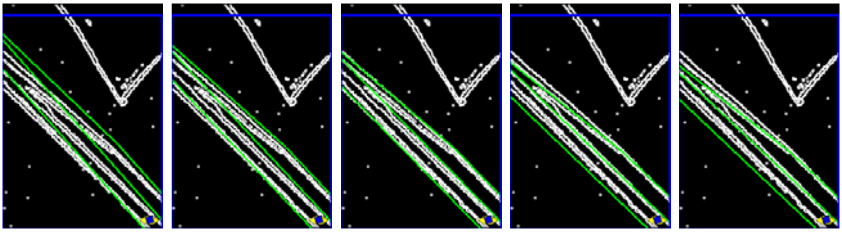


Fig. 4. GLDS: left gate leaf identification. A performance index is computed for each angular position; the estimated angular position is the one with the best index.

It needs to be noted that we compare the left gate leaf model with the pixels coming only from the left image and the same for the right one; we perform these operations only if a correct position in the last frame was found. The last condition avoids wrong behaviors, like following wrong objects or an extra enlargement of the detection window. The verification of the correctness of the estimated angular position is based on the following assumptions:

- **Quasi-constant gate leafs' velocity**: the difference between the actual velocity and the velocities detected in the past frames must be lower than a given threshold.
- **Movement consistency**: both gate leafs must perform the same movement (opening, closing, stopped).
- **Gate leaf priority**: one of the gate leaf is always the first that opens and the last that closes, showing a relationship between the relative position of the two gate leafs.

In order to use these assumptions we compute the velocity of a gate leaf as the angle covered during the acquisition of an image frame, and the velocity error as the difference between the current velocity value and the average of the velocities computed over the last frames. If the estimated angular position verifies all three assumptions, their value is considered correct and the algorithm continue, otherwise an error signal will be sent to the gate control board.

2.3 Obstacle Detection

In our application two obstacle detection (OD) algorithms are used: background subtraction based OD and stereo based OD. The background subtraction based OD that we implemented is a simple algorithm that works on grayscale images and uses a basic motion detection[2] concept. The algorithm is based on the well-known assumption that a moving object is made of colors which differ from those in the background and can be summarized by the following formula:

$$R_i = \begin{cases} 255, & d_i > t \\ 0, & otherwise \end{cases} \tag{1}$$

where R_i is the *i-th* pixel of the resulting image and d_i the difference of the same pixel in image I and background image B, as expressed in this formula:

$$d_i = |I_i - B_i| \tag{2}$$

This algorithm takes as input the images coming from the virtual cameras called *LC* and *RC* and produces grayscale images containing the obstacles in the scene; the background image is updated only when the gate is closed and *SC1* and *SC2* do not retrieve any obstacles in the surveillance area.

The Stereo OD algorithm works on the images provided by *SC1* and *SC2* virtual cameras; starting from these IPM images, a difference image D is generated comparing every pixel i of the left image to its homologous pixel of the right one and computing their distance as in the background subtraction. The resulting image D is then filtered with a particular low pass filter that can be summarized by the following equation:

$$\forall i \in D, \quad m = \frac{\sum_{\forall j \in A} D_j}{N_A} \quad T_i = \begin{cases} 0 & m < y \\ 1 & m > y \end{cases} \tag{3}$$

where N_A represents the number of pixels in A.

This kind of low pass filtering is useful to find the main differences in these images and is faster than other similar methods[1]. The images coming from both OD algorithms are then labelled: each connected area is localized and labeled with a progressive number for further identification and filtering. The main difference between the two OD systems is that the stereo one is able to retrieve other additional information about the obstacles, like their position in the danger area, their size and their height; these information are useful for filtering and tracking making the stereo OD results more robust.

3 Results

The test activity was focused on the main differences between our new safety sensor and the standard safety systems for automatic gate control. To do this, several video sequences of a real gate system was taken; then these video sequences were processed offline by our algorithm. We created 5 tests, each one representing strengths or weaknesses of the traditional safety photodetector and for all of these tests we have then evaluated the behavior of our new system. The tests are structured as follow:

* Opening and closing without obstacles.
* Presence of obstacles detectable by photodetectors.
* Presence of obstacles not detectable by photodetectors because of their size.
* Presence of obstacles not detectable by photodetectors because of their position.

The aim of the first test was to evaluate the precision of the gate leaf detection algorithm and to check the robustness of the OD algorithms; the results were very satisfactory, the *GLDS* showed high precision and the OD assured good robustness, as shown in fig.5 where the output of the gate leaf detection system is highlighted by the green lines that compose the gate model. The second test was created to evaluate the effectiveness of our sensor when compared with to the traditional security sensor; our sensor outperformed the original setup because the Stereo OD performs a very precise obstacle detection. The third test covers a case in which the traditional photodetector fails: as shown in fig.6 the objects, in this case represented by some packages, were too small to be detected by the photodetector. Our safety sensor detects these objects in the correct position (as showed by the red marker of fig.6) regardless of their location or size. In fig.6, as in all figures shown in this section, the algorithm's output is represented

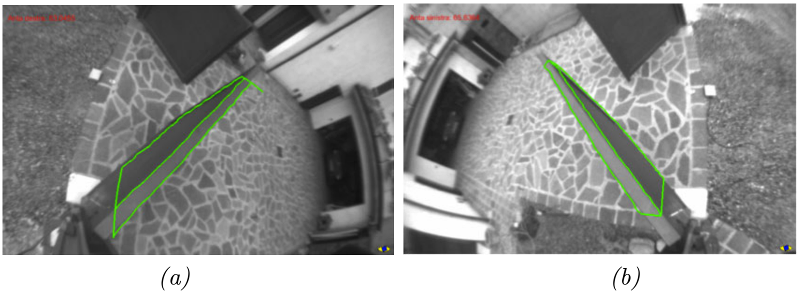

(a) (b)

Fig. 5. Left (a) and Right (b) gate leaf identification

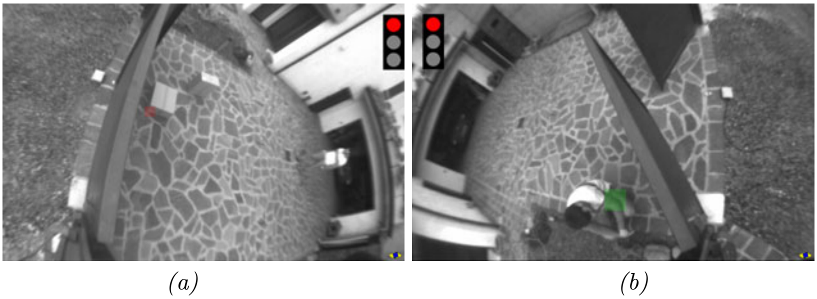

(a) (b)

Fig. 6. Obstacles Detection: small size obstacles (a) and obstacles hidden behind a gate leaf (b)

by a traffic light that become green if the gate leafs can move or turns to red if they would hit an obstacle. The fourth test covers another case in which photodetectors do not work: in this case the obstacles location were such that they were not visible from the photodetectors. This test was also very useful to verify the strength of our background subtraction algorithm: obstacles were often visible only by LC or RC, so they were visible only by the Background Subtraction based OD. Like in the other case, the results were extremely positive: the background subtraction algorithm provided enough robustness even with fast changing background and moving obstacles, as shown in fig.6 with a person hidden behind a gate leaf.

4 Conclusion and Future Works

In section 3 we saw that our advanced safety sensor showed satisfactory performance during the test stage: the sensor is able to detect obstacles in the whole danger area using different approaches for different zones of the danger area; furthermore the sensor showed an interesting robustness at the typical problems of outdoor vision system, as fast light changes, vibrations, wind and dazzle due to reflected sunlight. Despite that there are some situations that have to be thoroughly tested: i.e rain, snow and fog may request little changes in the obstacle detection algorithm and will be studied during the next months. Another interesting situation that we will take in to account in order to evaluate the system performances is the night scenario: neither the $GLDS$ nor the OD algorithm can work without natural sunlight, and the use FIR cameras must be excluded because of their cost. So, an artificial illumination system must be considered. The sensor described in this paper has patent pending.

References

1. Broggi, A., Medici, P., Porta, P.: StereoBox: a Robust and Efficient Solution for Automotive Short Range Obstacle Detection. EURASIP Journal on Embedded Systems (June 2007)
2. Benezeth, Y., et al.: Review and evaluation of commonly-implemented background subtraction algorithms. In: Pattern Recognition, ICPR (2008)
3. Bertozzi, M., Broggi, A., Fascioli, A.: Stereo inverse perspec- tive mapping: theory and applications. Image and Vision Computing 16(8), 585–590 (1998)
4. Collins, R.T., Lipton, A.J., Kanade, T.: A system for video surveillance and monitoring. Carnegie Mellon Univ., Pittsburgh (2000)
5. Bertozzi, M., Bombini, L., Broggi, A., Cerri, P., Grisleri, P., Zani, P.: GOLD: A Complete Framework for Developing Artificial Vision Applications for Intelligent Vehicles. IEEE Intelligent Systems (2008)
6. Informations on safe doors/gates, EN 12445 EN 12453

Using Blood Vessels Location Information in Optic Disk Segmentation

A.S. Semashko[1], A.S. Krylov[1], and A.S. Rodin[2]

[1] Laboratory of Mathematical Methods of Image Processing,
Faculty of Computational Mathematics and Cybernetics,
Lomonosov Moscow State University
http://imaging.cs.msu.ru

[2] Ophthalmology Chair, Faculty of Fundamental Medicine,
Lomonosov Moscow State University

Abstract. In this paper we present an approach to achieve high accuracy of optic disk segmentation using information on the location of blood vessels (*vessel map*). Morphological preprocessing is employed to remove the vessels from the image and to compute the vessel map. Vessel map is combined with edge map to obtain robust initial approximation of OD boundary using circular Hough transform. We use this approximation to build 2D weight function for the edge map, which is then used in the active contour model. We introduce an additional step to perform correction of the contour; in this step, the active contour model includes pressure forces and soft elliptical constraint. Vessel map is used in calculation of the ellipse parameters and pressure values. The method was tested on 1240 publicly available retinal images, and manual labeling of the disk boundary by medical experts was used to assess its accuracy and compare it with other optic disk segmentation methods.

1 Introduction

Optic disk (OD) is the region where optic nerve and blood vessels pass through the sclera. In automated retinal image analysis, locating the optic disk boundary is one of the most important tasks. Changes in optic disk appearance are used by ophthalmologists to assess the progression of diseases and treatment. Particularly, size of the optic disk cup is crucial in the study of patients with glaucoma. Other than being an indicator for various pathologies, optic disk is used as a reference to measure distances and detect other anatomical parts of the retina.

There are various methods for finding the exact shape of the disk. A thorough review is available in [17], so in our paper we will focus on the most referenced and/or close to our work methods only. Lalonde et al. [8] proposed an OD localization scheme using Hausdorff based template matching and pyramidal decomposition. This method assumes a circular model to approximate OD region for which radius parameter was estimated from the localized region. Li et al. [9] used an active shape model with some modifications that improve the

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 384–393, 2011.

method robustness in optic disk segmentation. Lowell et al. [10] used an elliptical deformable model with multi-stage boundary detection: the more salient temporal-side edge is detected first. In [4] a method based on blob detection is presented that is designed to be insensitive to image quality and not dependent on the image registration method.

Many OD segmentation methods employ specific models of disk shape, which makes them unable to precisely segment optic disks of irregular shapes. A number of freeform active contour based methods was presented recently to tackle irregular optic disk shapes as well. Mendels et al. [11] proposed to use gradient vector flow snake with a preprocessing step that removes blood vessels. Osareh et al. [13] improve the previous work by using a template matching scheme for automatic snake initialization and performing preprocessing in three-dimensional CIELab color space. Xu et al. [19] developed a freeform snake approach and report performance improvement over their previous results [9], however the segmentation accuracy is highly dependent on contour initialization. Tang et al. [16] proposed a Chan and Vese (C-V) model [2] based method, but their model incorporate an elliptical shape restraint. Joshi et al. [6] presented a C-V based method which does not impose any shape constraint to the underlying model. Instead, the model differentiates the OD region from the similar characteristic regions around it by integrating information from multiple image feature channels: the red color image channel and two texture-space channels.

Most freeform model-based methods include a preprocessing step that removes blood vessels, because their edges would affect the model evolution, severely distracting the result. In [6] blood vessels are segmented first, then the values at vessel points are somehow interpolated using nearby non-vessel regions. The approach employed in [11][13] uses, on the contrary, a simple but powerful vessel removal technique based on mathematical morphology, which works fine just because of the nature of the retinal image.

Optic disk segmentation methods based on some kind of freeform model that requires an image without blood vessels to perform well, share one common drawback: optic disk boundary in the places where it is crossed by blood vessels is not treated specifically. Such behavior implies that the vessel removal technique should recover the true optic disk shape, which is typically not the case. For example, the approach used in [11][13] often results in small concavities in the optic disk boundary in the places where it is crossed by blood vessels after the preprocessing.

In this paper we present a method based on [13]. In the preprocessing step we detect image areas occupied by blood vessels. This information is used later, both in the preprocessing and correction steps, resulting in significantly higher segmentation accuracy in most cases.

Notes. 1. The paper focuses on the segmentation and does not cover optic disk localization problem. 2. After the disk location is determined, we use only a part of the original image containing the disk to increase the performance. All operations described in this paper use only the cropped image.

2 Preprocessing

Mendels et al. [11] proposed to use grey-level mathematical morphology for removing blood vessels. Osareh et al. [13] carried out a series of experiments comparing the results of morphological closing operation in several color spaces, namely: grey-level, lightness channel L of HLS space, full HLS space, LCh space and Lab space. Between all the color spaces tested, the Lab space allows to get the most homogeneous region and preserves optic disk edges better.

Application of morphological operations in multidimensional color spaces requires a rule for comparing different color values. Osareh et al. used a simple definition of maximum and minimum in Lab space to perform morphological closing: to compare two color values, the Euclidean distance from the color space origin $(0,0,0)$ is computed for both colors and the resulting values are compared. This strategy works fine for optic disk images, though it might not perform well in other applications.

The difference between original image and the result of morphological closing I_c is used to calculate *vessel map* D. Each value in this map can be thought of as the Bayesian probability of the corresponding pixel being a vessel pixel. In rare cases a better segmentation may be achieved by applying morphological dilation with a small (1-2 pixels radius) disk-shaped structuring element to the vessel map. An example of a morphological closing result and the corresponding vessel map is shown in Fig. 1 (b, c).

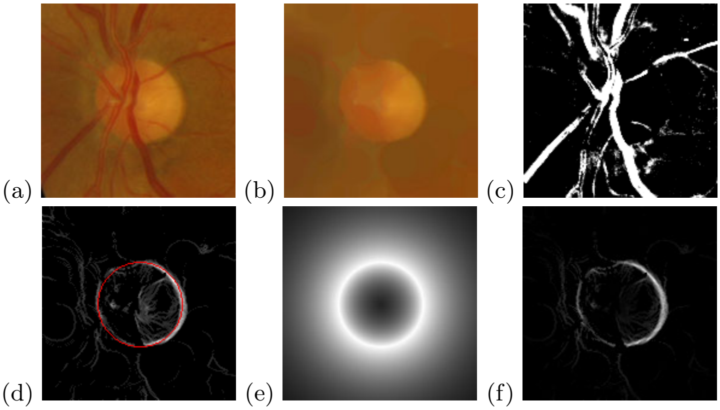


Fig. 1. Sample intermediate results of preprocessing stage. a) Original sub-image; b) result of blood vessels removal I_c; c) vessel map D; d) preliminary edge map M_{pre}, and the circle obtained using Hough transform (shown in red); e) edge map scaling coefficient; f) final edge map M.

The vessel map and the lightness channel L_c in I_c are used to produce a preliminary edge map as follows: $M_{pre}(P) = |\nabla L_c(P)| \cdot (1 - D(P))/\max|\nabla L_c|$. The Hough transform for circles is applied to the preliminary edge map to obtain a circular approximation of the optic disk boundary (see Fig. 1, d). This circle is used to construct an initial contour for the active contour model, and also to generate the final edge map M. The preliminary edge map typically contains a lot of edges that do not belong to the OD boundary. There may be some pathologies in close vicinity of the disk, and the OD itself often have highly inhomogeneous lightness, which results in strong edges inside the disk. These unwanted edges, though located at some distance of the initializer, may affect the active contour because of the nature of the GVF external force field that used in our work. To suppress them, we multiply each value in the preliminary edge map with a coefficient that is derived from the distance between the corresponding pixel and the circular approximation (see Fig. 1, e).

The resulting edge map M is used to compute the GVF external force field for the active contour model.

3 Active Contour Models

Active contour model, specifically the so-called *gradient vector flow* (GVF) snake [18], is used to localize the optic boundary, and the same model with pressure forces added to it is used in the correction process. Active contour is a parametric curve $\mathbf{v}(s) = (x(s), y(s))^{\mathrm{T}}$ ($s \in [0,1]$) that iteratively moves due to influence of internal and external (image and/or constraint) forces acting upon it. The original snake model [7] is formulated as an energy minimization problem. The basic form (that does not include external constraints) of the energy functional is as follows:

$$E[\boldsymbol{v}(s)] = \int\limits_0^1 \Big(\alpha(s)|\boldsymbol{v}(s)'_s(s)|^2 + \beta(s)|\boldsymbol{v}(s)''_{ss}(s)|^2 + E_{\mathrm{ext}}(\boldsymbol{v}(s))\Big)ds, \qquad (1)$$

where E_{ext} is determined by the edge map. For example, if edge strength is defined as squared magnitude of the intensity image I gradient, then image energy at point P is $E_{\mathrm{ext}}(P) = -|\nabla I(P)|^2$. In discrete formulation, Euler equations for the curve that minimizes Functional (1) are

$$\boldsymbol{A}\boldsymbol{x} + \boldsymbol{f_x}(\boldsymbol{x}, \boldsymbol{y}) = 0, \quad \boldsymbol{A}\boldsymbol{y} + \boldsymbol{f_y}(\boldsymbol{x}, \boldsymbol{y}) = 0, \qquad (2)$$

where A is a pentadiagonal banded matrix that depends only on the values of α, β and the spatial discretization step Δs. Here

$$\begin{aligned} \boldsymbol{f_x}(\boldsymbol{x}, \boldsymbol{y}) &= (f_x(x_1, y_1), f_x(x_2, y_2), \ldots, f_x(x_n, y_n))^{\mathrm{T}}, \\ \boldsymbol{f_y}(\boldsymbol{x}, \boldsymbol{y}) &= (f_y(x_1, y_1), f_y(x_2, y_2), \ldots, f_y(x_n, y_n))^{\mathrm{T}}, \\ f_x(x, y) &= \partial E_{\mathrm{ext}}(x, y)/\partial x, \\ f_y(x, y) &= \partial E_{\mathrm{ext}}(x, y)/\partial y, \\ \boldsymbol{x} &= (x_1, x_2, \ldots, x_n)^{\mathrm{T}}, \\ \boldsymbol{y} &= (y_1, y_2, \ldots, y_n)^{\mathrm{T}}. \end{aligned}$$

Equations (2) can be thought of as force balance equations; they are satisfied when the snake reaches equilibrium. The snake movement is governed by the following equations:

$$\begin{aligned} \boldsymbol{x}_t &= (\boldsymbol{A} + \Delta t\boldsymbol{I})(\boldsymbol{x}_{t-1} - \boldsymbol{f}_{\boldsymbol{x}}(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})), \\ \boldsymbol{y}_t &= (\boldsymbol{A} + \Delta t\boldsymbol{I})(\boldsymbol{y}_{t-1} - \boldsymbol{f}_{\boldsymbol{y}}(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})). \end{aligned} \tag{3}$$

The force balance Equations (2), and therefore snake evolution Equations (3), do not depend explicitly on the image energy E_{ext}. This allows to use more general definition of external forces, which is massively used in various extensions of the active contour model. In the original model, external forces have two serious drawbacks. First, they have low *capture range* of the edges, that results in the necessity of a very good contour initialization; when initialized far from the desired location, the active contour tends to converge to wrong result because often there are many local minima. The second drawback is that traditional active contour can not adequately detect boundary concavities (especially, the salient ones).

As mentioned before, the actual model used in this paper is based on the GVF snake [18]. It is one of the approaches addressing both drawbacks of the traditional model. It redefines image forces in a more complex way. Given an edge map $M(P)$ (for example, $M(P) = |\nabla I(P)|^2$), GVF force field $\mathbf{f}(x, y) = (f_x(x, y), f_y(x, y))$ is defined to be the field that minimizes the functional

$$\varepsilon[\mathbf{f}] = \iint \left(\mu\left(|\nabla f_x|^2 + |\nabla f_y|^2\right) + |\nabla M|^2(|\mathbf{f} - \nabla M|^2)\right) dxdy,$$

where μ is the regularization parameter that should be set according to the image noise level. In this formulation, the force field retains its smoothness in homogeneous regions, so the edge capture range is kept nearly as large as it is possible.

The correction step incorporates an additional force that attracts the contour to an ellipse. This force acts in the direction normal to the contour, and its magnitude at a control point is proportional to the distance between the point and the ellipse, bounded above by a constant.

Also, *pressure forces* [3] are used in the correction step. The idea is quite simple: each control point is additionally influenced by the pressure force that acts in normal direction either inwards or outwards.

4 Obtaining the Optic Disk Boundary

As mentioned in Sect. 2, the source image is morphologically closed, and the result is used to compute the edge map that defines the GVF force field, which is used in both active contour application steps (the main step and the correction step). In order to decrease processing time, the main step is split in two stages. The first one provides a rough result using large control point intervals and weak termination criteria; at the second stage, smaller control point interval is used and termination criteria are stronger.

When external force is applied to a control point, it is projected onto the normal to the contour at that point. This prevents movement of control points along the contour, so the difference between model states at different iterations can be expressed in terms of distances between single control point positions at these iterations. We use an inner iteration cycle (which typically consists of $t_{max} = 25$ iterations). Inside this cycle contour points are not redistributed, and a history of contour evolution is stored. Let $\boldsymbol{v}_{i,t}$ $(1 \leqslant i \leqslant n)$ be the i-th control point of the contour at t-th iteration of the inner cycle, then the termination criteria are:

$$\begin{cases} \max\limits_{t \in T} \max\limits_{i:\ 1 \leqslant i \leqslant n} u_{i,t} < c_1, \\ \max\limits_{t \in T} \frac{1}{n} \sum\limits_{i=1}^{n} u_{i,t} < c_2, \end{cases}$$

where $u_{i,t} = |\boldsymbol{v}_{i,t} - \boldsymbol{v}_{i,t_{max}}|^2$, $T = \{t_1, t_1+1, \ldots, t_{max}-1\}$, c_1 and c_2 are constants that define the required accuracy depending on the step being carried out, and $t_1 = 4$ is used to discard the effect of contour relaxation after its redistribution, which may cause significant changes to control point positions.

The correction step incorporates additional ellipse fitting and pressure forces. The intermediate contour is approximated by an ellipse using a fitting procedure that minimizes the sum of the squares of the distances, where each control point $\boldsymbol{v}_i$ have its weight equal to $1 - D(\boldsymbol{v}_i)$. The maximal ellipse fitting force magnitude is set to a relatively low value so that this force does not supersede image forces in the presence of a strong edge.

The pressure force slightly pushes the contour outwards in the regions where it is crossed by blood vessels. The magnitude of pressure force for k-th control point and the smoothness coefficient β_k depend on the corresponding values in vessel map D. Let

$$d_k = \frac{1}{\Delta s} \int\limits_{(k-1)\Delta s}^{(k+1)\Delta s} \left(1 - \frac{|s - k\Delta s|}{\Delta s}\right) \cdot D(\boldsymbol{v}(s)) ds,$$

then $\beta_k = (1 - d_k)\beta_{\text{normal}} + d_k \beta_{\text{vessel}}$, β_{normal} is equal to the value used in the second stage of the previous step, β_{vessel} is the value meant to impose *very strong* smoothness constraint, and $p_k^* = P_{\max}(1 - d_k)$, where p_k^* is a preliminary value for the pressure force magnitude, and $P_{\max} = 0.15$ is a constant denoting maximal pressure force value.

We also define two additional constants: threshold $P_{\text{thr}} = 0.1$ and a lowered pressure force value $P_{\text{low}} = P_{\max}/2$. The preliminary values vector $\mathbf{p}^*$ is convolved with Gaussian G_σ, where value of σ is selected so that the corresponding contour part length (assuming that control points are distributed uniformly) is equal to 10 pixels. Let $\tilde{\mathbf{p}}^* = \mathbf{p}^* * G_\sigma$ (note that $\mathbf{p}^*$ must be treated as discrete *periodic* function), then final pressure force magnitude values p_k are selected as follows:

$$p_k = \begin{cases} p_k^* & \text{if } \tilde{p}_k^* \leqslant P_{\text{thr}}, \\ \min(p_k^*, P_{\text{low}}) & \text{if } \tilde{p}_k^* > P_{\text{thr}}. \end{cases}$$

The aim of this procedure is to minimize the chance of excessive inflation of the contour. When there are areas with very dense blood vessels, the pressure is applied to a large number of adjacent contour points, making it possible for the contour to move away from the optic disk boundary. Using a lowered pressure force in such areas restores the balance.

The active contour receives new β_k values and is evaluated with pressure forces applied to it. In most cases this step requires very few iterations. When the contour reaches equilibrium, it becomes the final result.

5 Results and Concluding Remarks

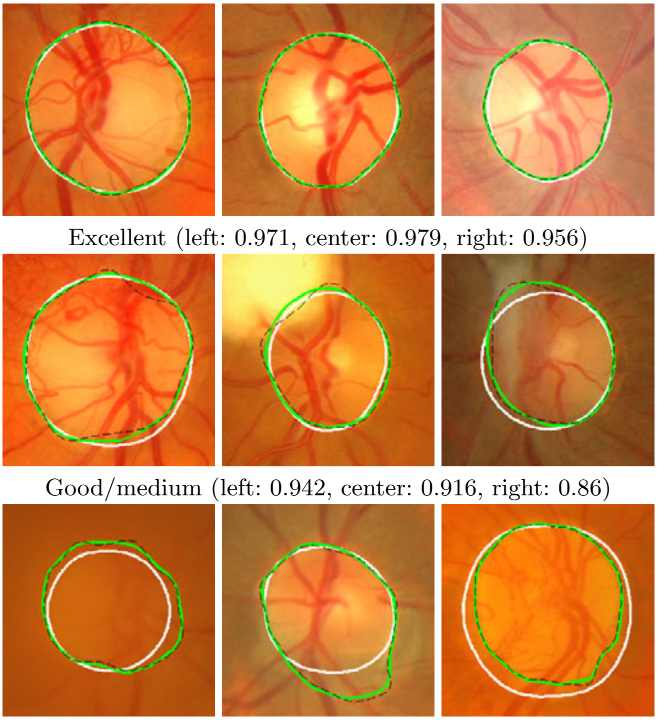
We tested our method on retinal images from two databases: MESSIDOR [12], kindly provided by the Messidor program partners, and DRIVE [15]. An expert ophthalmologist provided us with ground-truth optic disk contours for DRIVE images. Hand OD segmentations for MESSIDOR images, which we used in our testing, are currently available at [5]. Sample results of different quality are shown in Fig. 2; the measured accuracy of these results is also given.

We adopt a natural method of measuring the result accuracy A, the overlapping of the result R and the ground-truth contour G: $A = S(R \cap G)/S(R \cup G)$, where S denotes area. The overlapping degree is also used in [1], [8]. Table 1 shows percentage of images for several accuracy intervals, obtained by the compared methods, along with the mean accuracy of each method. In [1] the results of testing on MESSIDOR images are presented in the same manner and are included in Table 1. We did not include the method of Lalonde et al. [8] in the comparison, since Aquino et al. [1] obtained significantly better results. Table 1 also contains results of the proposed method for DRIVE images and results of the simple method based on GVF snake with color morphology preprocessing for both databases.

Table 1. Accuracy comparison using the contour overlapping. M and D denote MESSIDOR and DRIVE databases respectively.

Method	DB	$A \geqslant 0.95$	$A \geqslant 0.9$	$A \geqslant 0.85$	$A \geqslant 0.8$	$A \geqslant 0.75$	$A \geqslant 0.7$	$\overline{A}$
[1]	M	7%	46%	73%	84%	90%	93%	0.86
GVF	M	10.7%	39.3%	52.4%	59.2%	66.8%	72.7%	0.77
Proposed	M	20.9%	68.4%	82.6%	88.6%	91.7%	93.7%	0.89
GVF	D	17.5%	40.0%	45.0%	50.0%	57.5%	62.5%	0.74
Proposed	D	30.0%	77.5%	77.5%	80.0%	82.5%	92.5%	0.89

Aquino et al. claim that their method produces much better results (compared to the deformable model [10]) in the cases where an excellent segmentation is impossible to obtain due to poor contrast and/or severe pathologies. In a comparison where segmentation results were divided into *excellent*, *good*, *fair* and



Excellent (left: 0.971, center: 0.979, right: 0.956)

Good/medium (left: 0.942, center: 0.916, right: 0.86)

Bad (left: 0.795, center: 0.817, right: 0.804)

Fig. 2. Examples of segmentation results of different accuracy obtained with our method. Ground truth is shown in white, method result is shown in green, and the black dashed line represents intermediate contour before the correction stage.

poor, their method was inferior to the method of Lowell et al. only in the percentage of excellent segmentations (40% versus 42%), while the percentage of good and fair results was significantly higher (39% vs. 31% and 18% vs. 10% respectively). Comparing the results of testing on the MESSIDOR database, Table 1 shows that the proposed method has higher rate of good and fair segmentations, as well as excellent ones, than [1]. The comparison between [1] and [10] gives us an opportunity to compare our method with the Lowell's (though not strictly because of different accuracy measures). It seems that their excellent segmentation quality roughly corresponds to overlapping greater than about 0.91. It is mentioned that they obtained excellent results in 42% of the cases, while our method achieves overlapping greater than 0.91 in 63.2% cases. It seems improbable that their method has a higher rate of very precise segmentations ($A \geqslant 0.95$), and it is clear that in more difficult cases our method performs much better.

OD segmentation accuracy is also often measured in terms of binary classification of the image pixels: true positive T_p, true negative T_n, false positive F_p and false negative F_n pixel numbers are counted, and then the accuracy is expressed in terms of true positive rate (*sensitivity*) $R_{TP} = T_p/(T_p + F_n)$, true negative rate (*specificity*) $R_{TN} = T_n/(T_n + F_p)$, *positive predictive value*, or *precision rate*, which is defined as $P = T_p/(T_p + F_p)$, and *F-measure* $F = 2 \cdot R_{TP} \cdot P/(R_{TP} + P)$. These values suit well in case of binary classification of independently treated objects, but are not very adequate when used to measure similarity between two contours. Note that contour overlapping, which can be expressed as $T_p/(T_p+F_p+F_n)$, is always not greater than R_{TP} and P, and R_{TN} does depend on the size of the image part that is used for OD segmentation: counting pixels of the entire image instead of a cropped part of it would result in very high specificity values, and these values will have little meaning.

We calculated these characteristics for our method and got the following average values: $R_{TP} = 94.1\%$, $R_{TN} = 98.4\%$, $P = 94.3\%$, $F = 0.94$. The implicit active contours-based method in [14] achieved 90.67% average sensitivity and 94.06% average specificity; a local database of 148 images was used for testing. Duanggate et al. [4] reported average R_{TP} and P values of 85.37% and 82.87% respectively on cropped images. It is noted that the testing was done using two datasets: one containing mostly images with clearly visible OD (123 images total), and the second where ODs were mostly faint and unclear (91 image). In [6], the method was tested on 138 images, resulting in the following average values: $R_{TP} = 96\%$, $P = 98\%$ and $F = 0.97$.

Apparently, the proposed method provides better segmentations than [14] and [4]. The method presented in [6] possibly performs better, but in order to make a strong conclusion a rigorous testing of the two methods is needed involving the same retinal image database, ground truth and accuracy metrics. Unfortunately, Joshi et al. did not provide the results of testing their method on publicly available retinal image databases. It is noted that the method was tested on high quality OD-centric images, while DRIVE and MESSIDOR databases contain a noticeable amount of images where it is hard to achieve good OD segmentation, i.e. blurred and/or low-contrast images, or images with the OD located near the edge of camera field of view and affected by edge flare.

Thus testing shows that the method presented in this paper outperforms many other OD segmentation methods and has a high rate of very good segmentations. A great increase in accuracy, comparing to a simple GVF snake algorithm with morphological preprocessing, is achieved by using the blood vessel map information in all processing steps.

Acknowledgments. The work was supported by the federal target program "Scientific and scientific-pedagogical personnel of innovative Russia in 2009-2013" and a Moscow government grant.

References

1. Aquino, A., Gegundez-Arias, M., Marin, D.: Detecting the Optic Disc Boundary in Digital Fundus Images Using Morphological, Edge Detection and Feature Extraction Techniques. IEEE Transactions on Medical Imaging 29, 1860–1869 (2010)
2. Chan, T., Vese, L.: Active contours without edges. IEEE Trans. Image Processing 10(2), 266–277 (2001)

3. Cohen, L.: On active contour models and balloons. CVGIP: Image Understanding 53(2), 211–218 (1991)
4. Duanggate, C., Uyyanonvara, B., Makhanov, S.S., Barman, S., Williamson, T.: Parameter-free optic disc detection. Computerized Medical Imaging and Graphics 35(1), 51–63 (2011)
5. Expert system for early automated detection of DR by analysis of digital retinal images project website. Univ. Huelva, Huelva, `http://www.uhu.es/retinopathy`
6. Joshi, G.D., Gautam, R., Sivaswamy, J., Krishnadas, S.R.: Robust optic disk segmentation from colour retinal images. In: Proceedings of the 7th Indian Conference on Computer Vision, Graphics and Image Processing, pp. 330–336 (2010)
7. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. International Journal of Computer Vision 1(4), 321–331 (1988)
8. Lalonde, M., Beaulieu, M., Gagnon, L.: Fast and robust optic disc detection using pyramidal decomposition and Hausdorff-based template matching. IEEE Trans. Med. Imaging 20(11), 1193–1200 (2001)
9. Li, H., Chutatape, O.: Boundary detection of optic disk by a modified ASM method. Pattern Recognition 36(9), 2093–2104 (2003)
10. Lowell, J., Hunter, A., Steel, D., Basu, A., Ryder, R., Fletcher, E., Kennedy, L.: Optic nerve head segmentation. IEEE Trans. Medical Imaging 23(2), 256–264 (2004)
11. Mendels, F., Heneghan, C., Thiran, J.: Identification of the optic disk boundary in retinal images using active contours. In: Proc. Irish Machine Vision and Image Processing Conference on Cerebrovascular Diseases, pp. 103–115. IEEE, Los Alamitos (1999)
12. MESSIDOR: Digital Retinal Images, MESSIDOR TECHNO-VISION Project, France, `http://messidor.crihan.fr/download-en.php`
13. Osareh, A., Mirmehdi, M., Thomas, B., Markham, R.: Colour morphology and snakes for optic disc localisation. In: 6th MIUA Conference, pp. 21–24 (2002)
14. Siddalingaswamy, P., Gopalakrishna, P.: Automatic Localization and Boundary Detection of Optic Disc Using Implicit Active Contours. International Journal of Computer Applications 1(6), 1–5 (2010)
15. Staal, J., Abramoff, M., Niemeijer, M., Viergever, M., van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. IEEE Trans. Medical Imaging 23(4), 501–509 (2004)
16. Tang, Y., Li, X., von Freyberg, A., Goch, G.: Automatic segmentation of the papilla in a fundus image based on the C-V model and a shape restraint. In: Proc. ICPR, pp. 183–186 (2006)
17. Winder, R.J., Morrow, P.J., McRitchie, I.N., Bailie, J.R., Hart, P.M.: Algorithms for digital image processing in diabetic retinopathy. Computerized Medical Imaging and Graphics 33(8), 608–622 (2009)
18. Xu, C., Prince, J.: Snakes, shapes, and gradient vector flow. IEEE Trans. Image Processing 7, 359–369 (1998)
19. Xu, J., Chutatape, O., Sung, E., Zheng, C., Chew, P.: Optic disk feature extraction via modified deformable model technique for glaucoma analysis. Pattern Recognition 40(7), 2063–2076 (2007)

Orthophotoplan Segmentation and Colorimetric Invariants for Roof Detection

Youssef El Merabet[1,2], Cyril Meurie[1], Yassine Ruichek[1], Abderrahmane Sbihi[3], and Rajaa Touahni[2]

[1] Systems and Transportation Laboratory, Université de Technologie de Belfort-Montbliard, 13 rue Ernest Thierry-Mieg, 90010 Belfort, France
[2] Laboratoire LASTID, Département de Physique, Faculté des Sciences, Université Ibn Tofail, B.P 133, 14000 Kénitra, Maroc
[3] ENSA, Université Abdelmalek Essadi, Route Ziaten, km 10, BP 1818, Tanger Maroc
{youssef.el-merabet,cyril.meurie,yassine.ruichek}@utbm.fr,
sbihi_abderrahmane@yahoo.fr, rtouahni@hotmail.com

Abstract. In this paper, we use a morphological segmentation method called watershed for segmenting roof of “orthophotoplan” images. This work takes place in a global approach which consists in recognizing a roof of aerial images among a knowledge database and bending out 3D models automatically generated from geographical data. The main aim of this work consists in defining the best couple of colorimetric invariant/gradient (among 24 colorimetric invariants and 14 gradients tested) used as input of watershed algorithm in order to obtain the best segmentation of roof. The tests are made on a database of 67 roofs containing a certain heterogeneity (illumination changes, shadows, etc) and evaluated with the Vinet criteria (including a ground truth image) in order to prove the robustness of the proposed strategy.

Keywords: watershed, color gradient, colorimetric invariant, orthophotoplan.

1 Introduction

The works presented in this paper appear in a global approach that consists in recognizing a roof of aerial images among a knowledge database and bending out 3D models automatically generated from geographical data. The first step of the global approach presented in this paper consists in segmenting the roof in different regions of interest in order to provide several measures of the roof (section of roof, chimneys, roof light, etc). To do that, we use a morphological segmentation method called watershed. But this method requires two input images (seed and potential images) that we have optimized for the application. A first difficulties consists in using an appropriate potential image (gradient image) in order to extract as well as possible the details of the roof (chimneys, roof light, etc). For that, several tests have been performed to define the best suitable

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 394–403, 2011.

gradient among a set of 14 gradients (8 gray levels gradients and 6 color gradients). A second difficulties concerns the recognition of the roof among a database including the same roof with different illumination changes and shadows. That is why, we propose in this paper to choose an appropriate colorimetric invariant among a set of 24 invariants extracted of the literature to limit these effects. All tests presented in this paper have been performed on an "orthophotoplan" image containing 67 roofs. Each roof to segment is extracted of the "orthophotoplan" image from the ground track (red border in Figure 1). Figure 1 illustrates an "orthophotoplan" image (left), a zoom of one roof to extract/recognize (middle) and the segmented image to obtain (right). The paper is organized as follows: Section 2 details the proposed approach including a recall of different invariants and the watershed algorithm. In section 3, we detail all tests which are permit to choose the optimal couple of invariant/gradient of the watershed algorithm for the given application. Finally we conclude and present perspective of future works.

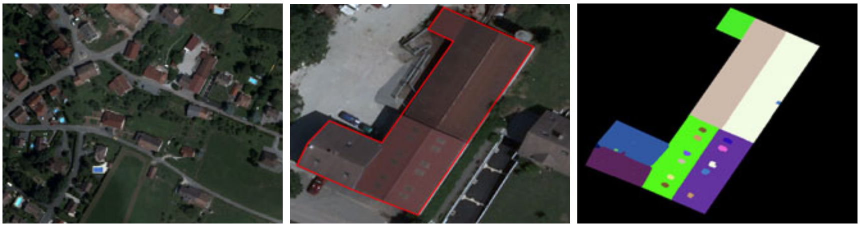


Fig. 1. Example of "orthophotoplan" image (left to right: an "orthophotoplan" image, a zoom of roof, the segmented image)

2 Image Segmentation

Image segmentation consists in partitioning an image in more or less regular or homogeneous regions according to a given criteria. Many segmentation methods exist in the literature. These one can be grouped in three categories: 1/ region based segmentation (split and merge, region growing in which we find the watershed algorithm used in this paper); 2/ edge based segmentation; 3/ classification, clustering, thresholding.

In this paper, we use the first method which corresponds to the region based segmentation and particulary a morphological method called color watershed that offers in general very good results. Thus, we will define the best couple invariant/gradient used as input images of the watershed algorithm.

Figure 2 illustrates the synopsis of the global proposed approach. It is composed of several steps. The first one corresponds to a colorimetric invariant to apply on the initial image. After that, we calculate a color gradient or a gray level gradient on the simplified image. In the case of the gray level gradient, we must extract the three component of the simplified image and calculate the

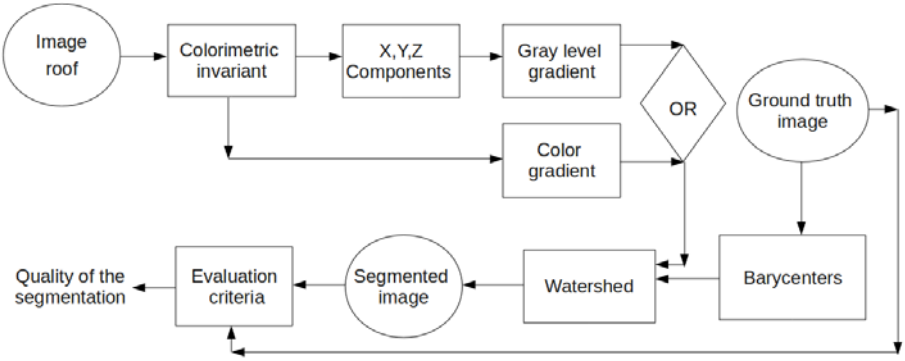


Fig. 2. Synopsis of the proposed approach

gradient on these component. The next step consists in using the watershed with the gradient image and a seed image (where a seed corresponds to the barycenter of each region of the ground truth image). Finally, we obtain the segmented image and evaluate the quality of the segmentation, with the Vinet criteria (and the reference segmentation).

2.1 Colorimetric Invariants

In our application, the "orthophotoplan" images contain a certain heterogeneity in terms of lights, illumination changes, shadows, etc. It does not permit to extract correctly the different regions of interest of the roof. To overcome these drawbacks, we rejoin the strategy of many authors by simplifying the input image with a suitable colorimetric invariant([1], [2], [3], [4], [5]).

Indeed, for few years, the color invariance generate much interest and continues to engage the field of computer vision. For example, one can cite the use of colorimetric invariant for matching images [4], for motion estimation in video sequences [5], for feature extraction and re-identification of individuals in transport environment [2], for enhancing the monitoring of points of interest in color images [3], etc.

In this paper, we want to show that using a colorimetric invariant can limit artefacts of the acquired image and thus obtain a better segmentation of the roof. That is why, we propose to define the best colorimetric invariant according to the proposed approach. 24 colorimetric invariants of the literature and listed below have been tested: Greyworld normalization (called Greyworld in Figures 4, 6) [6], RGB-rang [7], affine normalization (called affine in Figures 4, 6) [8], intensity normalization (called chromaticity in Figures 4, 6) [10], comprehensive color normalization (called comprehensive in Figures 4, 6) [6], c1c2c3 ([4],[1]), m1m2m3 [4], l1l2l3 ([4],[1]), l4l5l6 [11], A1A2A3 [3], c4c5c6 [11], hsl , MaxRGB [10], CrCgCb [3], Color Constant Color Indexing (called CCCI in Figures 4, 6) [10], m4m5m6 [3], Standard L2 (called L2 in Figures 4, 6) [3], Maximum-intensity normalization (called Mintensity in Figures 4, 6) [12], reduced coordinates [9],

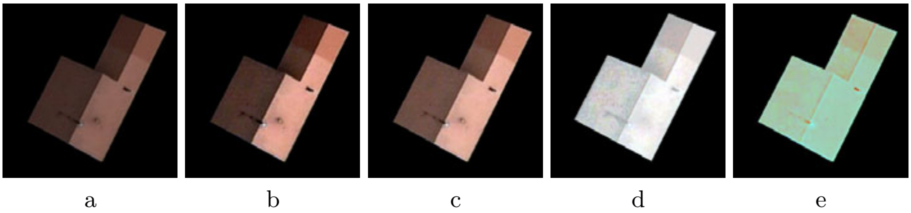


Fig. 3. Example of colorimetric invariants, (a) without colorimetric invariant, (b) with affine normalization, (c) with Maximum-intensity normalization, (d) with RGB-rang, (e) with c1c2c3

CrCb ([3],[1]), opposite colors (o1o2) ([3],[1]), Saturation S [4], Log-Hue [10] and Hue H ([4],[9]).

Figure 3 illustrates the influence of 4 colorimetric invariants applied on the initial image.

2.2 Watershed Algorithm

As indicated previously, we have chosen to use a morphological segmentation method called watershed which offers generally good result. Many versions of this algorithm exist in the literature ([13], [14] [15]), but we have chosen the version presented by Meyer ([14], [16]) that we recalled below :

This algorithm extends as soon as possible, the local minima of the image (in our case: the seed calculated on the reference image) using the priority given by a potential or gradient image. It is composed of 4 steps (algorithm 1).

The watershed algorithm requires two input images (a seed and a gradient/potential images), that we detail below:

The gradient or potential image: In order to define the best gradient for our application, eight gray level gradients and six color gradients have been tested. The eight gray level gradients are: 1/Sobel; 2/Roberts; 3/Prewitt; 4/GradientF (the first derivative of the image); 5/NonMaximaSuppression (the non maxima values from the magnitude of the gradient); 6/Shen; 7/Deriche; 8/GradientM (morphological gradient corresponding to the substraction between dilated image and eroded image). The six color gradients are: 1/Di-Zenzo [17]; 2/GradientMC (morphological gradient corresponding to the substraction between dilation and erosion using a lexicographical order); 3/GradientC (the marginal gradient) [18]; 4/SobelC (Sobel calculated on color image); 5/SobelTLS (Sobel calculated in TSL color space); 6/Carron [19].

The seed image: In the watershed algorithm, the seed image corresponds generally to the local minima of the gradient image. But, it leads to an over-segmentation of the image. To overcome this drawback, a solution consists to use a selection of local minima such as defined in [20] and which offers generally good results. But in order to define in the best conditions, the couple of invariant/gradient which offers the best segmentation results, we have chosen to use the knowledge of the application. In fact, we calculate the barycenter of each region of the ground truth image and consider that it correspond to a seed.

Algorithm 1. Watershed algorithm defined by Meyer ([14], [16])

```
begin
    1. Assign a label to each seed of image. Initialize a set S (type of hierarchical queue) to the empty set.
    2. Insert each labeled point in the set S. At each insertion, the queue established a tri of points by priority according to their altitude (ie module gradient).
    3. Extract a point x of set S with minimum altitude (low gradient) ie, $F(x) = min\{F(y)|y \in S\}$ (where F is the initial image). Assign each point y adjacent to x (in a neighborhood $\mathcal{V}$) and non-labeled, the label of x, and insert y in S.
    4. Repeat step 3 until S is not empty.
end
```

3 Experimental Results

To compare the segmentation results obtained with different gradients and colorimetric invariants, an appropriate evaluation is then necessary. Many evaluation methods exist in the literature. These can be classified into two categories: without and with ground truth (reference segmentation). In our application, we have used the Vinet criteria [21] which belongs to the second category. The evaluation of the segmentation results presented in this paper has been made on one "orthophotoplan" image containing 67 roofs. Each roof of the "orthophotoplan" image is associated to a ground truth (reference segmentation) created manually by a human expert. For a better readability, we have only presented eighteen better invariants among the 24 tested. As indicated in the section 2, the proposed approach uses the watershed algorithm (with an appropriate seed and potential images) on a simplified image (with a colorimetric invariant).

To define the best couple colorimetric invariant/gradient, we separate all tests in two categories. The first one corresponds to the best couple color gradient/colorimetric invariant applied directly on the simplified image (cf Figure 6). The second one corresponds to the best couple gray level gradient/component of the colorimetric invariant applied on the component of the simplified image (cf Figure 4). In this last case, X, Y and Z represents respectively the first, second and third component of the colorimetric invariant. For example, hsl-Y corresponds to the second component: the saturation (S).

Figures 4 and 6 illustrate respectively the best couple component (of the colorimetric invariant)/gray level gradient and the best couple colorimetric invariant/color gradient. For both figures, each bar represents the mean value of Vinet calculated on all images of database according to the couple invariant/gradient

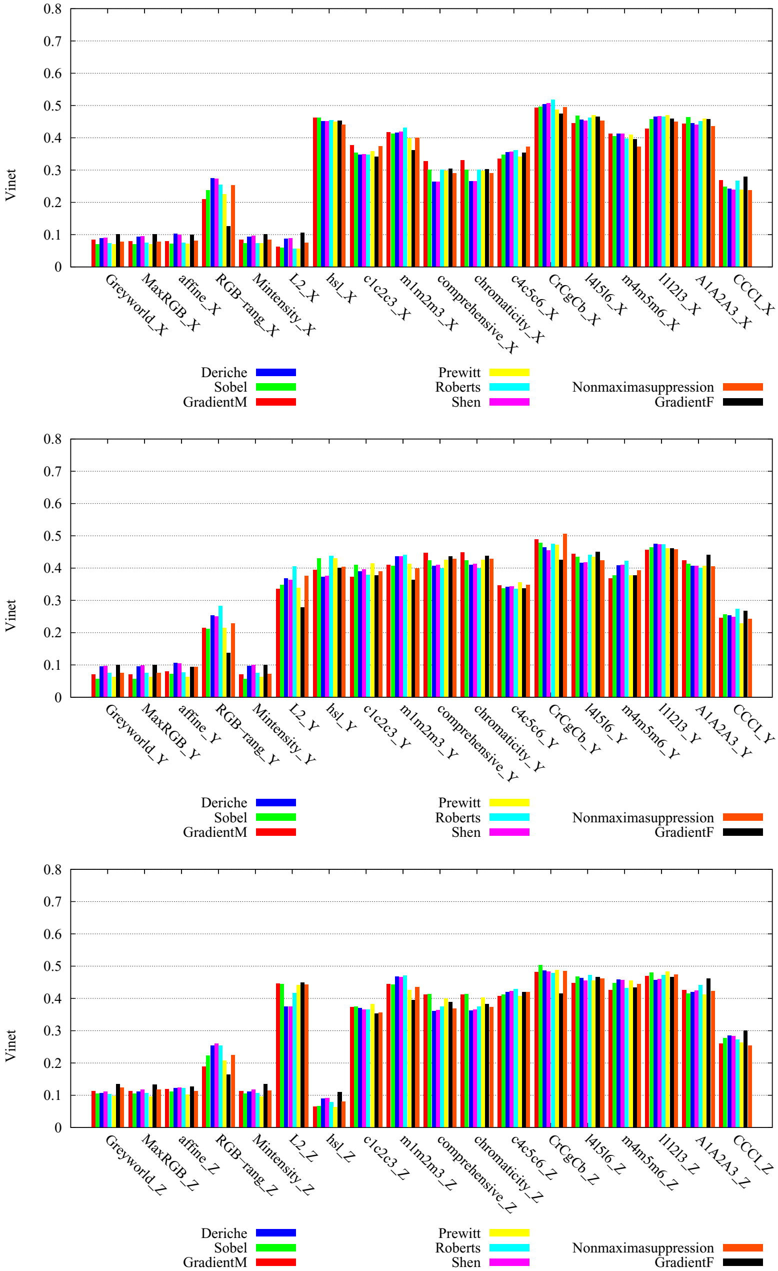


Fig. 4. Quality of the segmentation on gray level image (using Vinet criteria) according to the couple colorimetric invariant/gradient. (top to bottom: segmentation results on the X component (top), Y component (middle), Z component (bottom) of the color image).

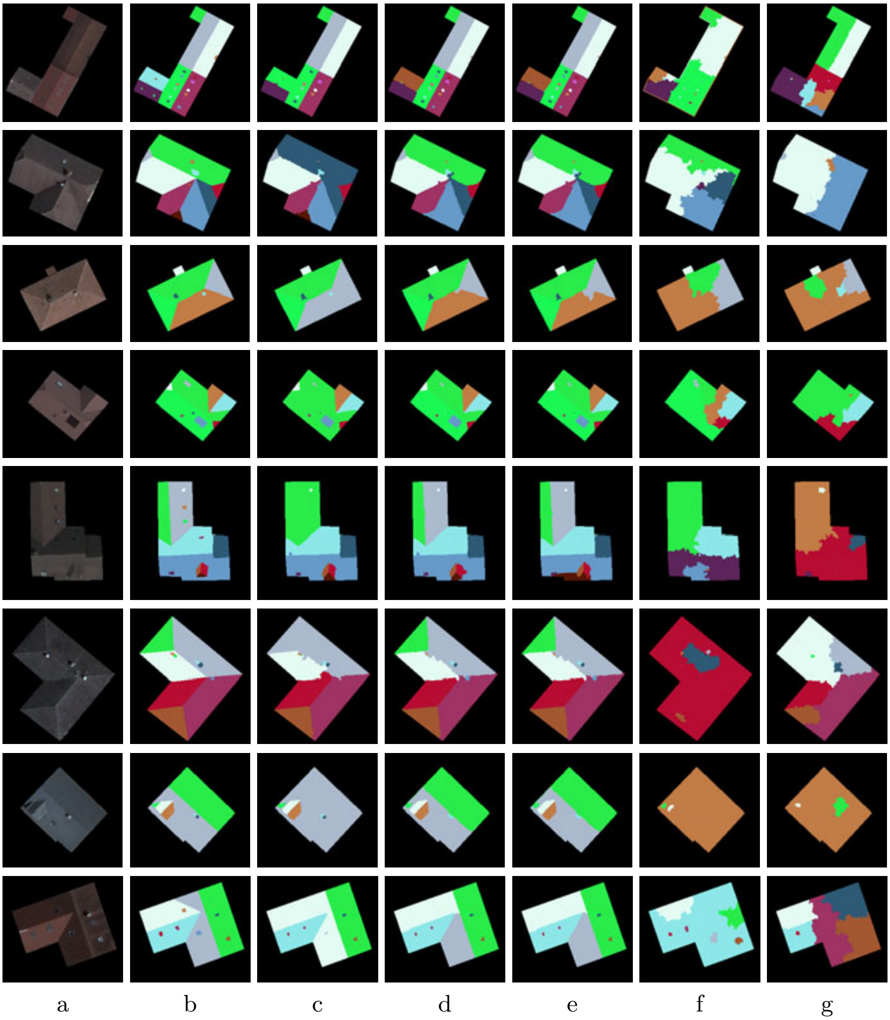


Fig. 5. Illustration of few segmented images obtained without/with colorimetric invariant. Initial image (a), ground truth image (b), segmented image without invariant (c), with Maximum-intensity normalization (d), with Max-RGB (e), with l1l2l3 (f) and with m1m2m3 (g).

tested (the results are better if the value is low). For example, considering the first component of Greyworld invariant (Greyworld-X), the first bar (in red color) corresponds to the Vinet value calculated with the GradientM gradient.

Considering the gradient calculated of the gray level image (cf Figure 4), the segmentation results depends highly of the color component selected. Indeed, for the three X, Y and Z components of the colorimetric invariants Greyworld, affine normalization, RGB-rang, Maximum-intensity normalization, and MaxRGB give satisfying results (the values of Vinet are low compared to those obtained with the components of the other colorimetric invariants) and whatever the gradient used. But, if we select only the best component, one can notice that the X component of the L2 normalization (L2-X) and the Z component of the hsl space (hsl-Z) give the best segmentation results. Finally, we can define a preference order to use with a couple of gray level gradient/component of colorimetric invariant: Prewitt/L2-X, Roberts/L2-X, GradientM/L2-X, Deriche/Mintensity-Y.

Considering the gradient calculated on the color image (cf Figure 6), Greyworld, affine normalization,RGB-rang, Maximum-intensity normalization and MaxRGB, give the best segmentation results, whatever the color gradient used even if the couple Di-Zenzo/Mintensity is the best. We can also define a preference order to use with a couple of color gradient/colorimetric invariant: Di-Zenzo/Mintensity, Di-Zenzo/MaxRGB, Di-Zenzo/Greyworld, Di-Zenzo/affine normalization, color morphological gradient/MaxRGB, GradientC/Mintensity. For a better visualization, the figure 5 illustrates few segmentation results obtained by the proposed approach. The Di-Zenzo gradient is used for this illustration since it permits to obtain the best results. One can notice that segmentations using an appropriate gradient/invariant offers better results than those obtained without colorimetric invariant. The segmentation results obtained with Maximum-intensity normalization and MaxRGB and presented on the figure 5 are satisfying and confirm the conclusion given about the figure 6. The segmentations obtained without and with l1l2l3 (cf Figure 5(f)), m1m2m3 (cf Figure 5(g)) colorimetric invariants appear weak. It is illustrating by a partial or total loss of roof information with worse quality of edge.

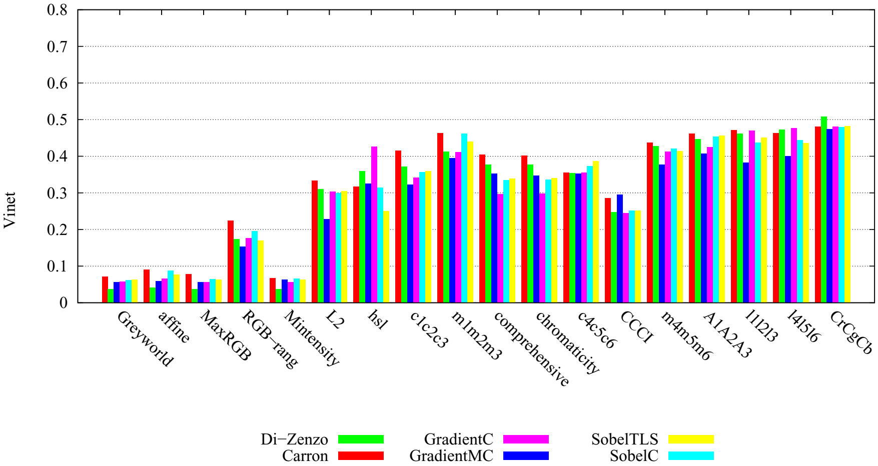


Fig. 6. Quality of the segmentation on color image (using Vinet criteria) according to the couple colorimetric invariant/gradient

4 Conclusion

An image segmentation method based on watershed algorithm using appropriate couple of colorimetric invariant/gradient is presented. This segmentation method is proposed in the field of “orthophotoplan” images segmentation for roof detection. The tests performed on database with 67 roofs show that the Di-Zenzo gradient coupled to the Maximum-intensity invariant gives the best results if we consider the color information. In opposition, if we consider the gray level information, the Prewitt gradient coupled with the X component of the L2 invariant gives the best results. The proposed approach permits to conclude on the importance of the colorimetric invariant and gradient used in the segmentation step for our application. Future works concern the extraction of roof measures in order to create automatically several types of 3D model of roof.

Acknowledgements. The authors would like to thank the Volubilis program for their financial support, and the “Communauté d'agglomération Belfortaine” for the “orthophotoplan” images on which the tests have been performed.

References

1. Gevers, T., Smeulders, A.: Object Recognition based on Photometric Colour Invariants. In: Proceedings of SCIA, Lappeenranta, Finland (1997)
2. Cong, T., Khoudour, L., Achard, C., Meurie, C., Lezoray, O.: People reidentification by spectral classification of silhouettes. Signal Processing 90(8), 2362–2374 (2010)
3. Gouiffès, M.: Apports de la Couleur et des Modéles de Rèflexion pour l'Extraction et le Suivi de Primitives, Thése de doctorat, Université Poitiers (Décembre 2005)
4. Gevers, T., Smeulders, A.: Colour based object recognition. Pattern Recognition 32, 453–464 (1999)
5. Golland, P., Bruckstein, A.M.: Motion from color. Computer Vision and Image Understading 68(3), 346–362 (1997)
6. Schaefer, G.: How useful are colour invariants for image retrieval. In: Computational Imaging and Vision, Proc. Int. Conference on Computer Vision and Graphics, Warsaw, Poland, vol. xx. Kluwer Academic Publishers, Dordrecht (2004)
7. Hordley, D., Finlayson, G.D., Schaefer, G., Tian, G.Y.: Illuminant and device invariant colour using histogram equalization. Pattern Recognition 28, 179–190 (2005)
8. Fusiello, A., Trucco, E., Tommasini, T., Roberto, V.: Improving feature tracking with robust statistics. Pattern Analysis & Applications 2(4), 312–320 (1999)
9. Gevers, T., Stockman, H.: Robust histogram construction from color invariants for object recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 26, 113–118 (2004)
10. Hordley, S.D., Finlayson, G.D., Schaefer, G., Tian, G.Y.: Illuminant and device invariant colour using histogram equalization. Elsevier Science, Amsterdam (2002)
11. Latecki, L.J., Rajagopal, V., Gross, A.: Image retrieval and reversible illumination normalization. In: Proc. of the IS&T/SPIE. Internet Imaging VI, San Jose (2005)
12. Dargham, J.A.: Lip detection by the use of neural networks. Artif. Life Robotics 12, 301–306 (2008)

13. Cousty, J.: Lignes de partage des eaux discrétes: théorie et application a la segmentation d'images cardiaques. PhD thesis, Université Marne-la-Vallé (2007)
14. Meyer, F.: Un algorithme optimal de ligne de partage des eaux. In: Dans Actes du 8éme Congrés AFCET, Lyon-Villeurbanne, France, pp. 847–859 (1991)
15. Vincent, L., Soille, P.: Watersheds in digital spaces. An efficient algorithm based on immersion simulations. IEEE Trans. Pattern Analysis and Machine Intelligence 13(6), 583–598 (1991)
16. Cousty, J.: Lignes de partage des eaux discrétes: théorie et application a la segmentation d'images cardiaques. PhD thesis, Université Marne-la-Vallé (2007)
17. Di Zenzo, R.: A note on the gradient of a multiimage. Computer Vision, Graphics and Image Processing 33, 116–125 (1986)
18. Lezoray, O., Elmoataz, A., Cardot, H., Revenu, M.: Segmentation d'images couleur: applications en microscopie cellulaire. Traitement du Signal 17, 33–45 (2007)
19. Carron, T.: Segmentation d'images couleur dans la base Teinte Luminance Saturation: approche numérique et symbolique. Thèse de doctorat, Thèse de l'Université Savoie soutenue (décembre 1995)
20. Cohen, A., Attia, D., Meurie, C., Ruichek, Y.: Une méthode de segmentation hybride par combinaison adaptative des informations texture et couleur. In: Conférence MAJESTIC, Bordeaux, France (2010)
21. Vinet, L.: Segmentation et mise en correspondance de régions de paires d'images stéréoscopiques. PhD thesis, Université Paris IX Dauphine, Juillet (1991)
22. Gevers, T., Stockman, H.: Classifying of color edges in video into shadow-geometry, highlight, or material transitions. IEEE Transactions on Multimedia 5(2), 237–243 (2003)

A Simulation Framework to Assess Pattern Matching Algorithms in a Space Mission

Alessandro Gherardi[1] and Alessandro Bevilacqua[1,2]

[1] ARCES – Advanced Research Center on Electronic Systems,
University of Bologna, Italy
agherardi@arces.unibo.it

[2] DEIS – Department of Electronics, Computer Science and Systems,
University of Bologna, Italy
alessandro.bevilacqua@unibo.it
http://cvg.deis.unibo.it

Abstract. Within the framework of the European Space Agency (ESA), the BepiColombo space mission will target Mercury as the planet to be studied to discover more about the formation and the composition of the inner planets of our solar system. Mercury exhibits an effect (the *libration*) whose amplitude could determine whether the core is liquid.

The method we propose to estimate the libration amplitude is based on matching of surface features extracted from image pairs taken at different epochs and thereby affected by illumination and scale artifacts. Since no detailed image pairs are available to assess the accuracy of the methods we propose, in order to generate synthetic images of the planet's surface we have developed the simulation framework we present in this work. Finally, we discuss some preliminary matching results on surface features extracted from our synthetic images taken at different altitude and affected by illumination changes.

Keywords: synthetic image, satellite imaging, pattern matching, simulation framework.

1 Introduction

The BepiColombo mission is the cornerstone mission of ESA that will target Mercury to discover the formation and the composition of the inner planets of our solar system. The main challenges arise from the planet being very close to the Sun, which affects many technological aspects. Firstly, Mercury is hard to observe from a distance because the Sun is very bright. Furthermore, it is difficult to reach because a spacecraft must loose a lot of energy to approach the planet and enter in its nominal orbit. Moreover, the high gravity field of the Sun presents a challenge in placing the spacecraft into a stable orbit around Mercury. Once in a stable orbit, the spacecraft must also cope with high thermal issue, since its surface temperatures may reach 450°C. The spacecraft, will be launched in 2014 and will have a interplanetary cruise to Mercury using solar-electric propulsion and by exploiting Moon, Venus and Mercury gravity assists.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 404–413, 2011.

The Mercury Planetary Orbiter (MPO) probe will be captured into polar orbit stabilized in nadir pointing, which will have periherm distance of 400 km and a apoherm distance of 1500 km, with 2.3 hours orbital period [1]. Among the instruments carried by MPO, an imaging system will provide detailed images of the surface of Mercury, through wide angle and narrow angle High Resolution Cameras (HRIC). One of the main objective of the mission is the Mercury Orbiter Radio science Experiment (MORE) [2]. The MORE experiment will help to determine the gravity field of Mercury, as well as the size and physical state of its core. It will provide crucial experimental constraints to the models of the planet's internal structure and test theories of gravity with unprecedented accuracy. The main objective of the rotation experiment is the estimation of Mercury obliquity and libration amplitude: the physical libration of the mantle about the mean resonant angular velocity arises from the periodically reversing torque on the planet as Mercury rotates relative to the Sun. Roughly, such an effect will be seen as an offset in longitude of the same point on the Mercury surface observed by the orbiting probe at different phases of the libration. The amplitude of this libration is approximately equal to 400m at the equator, following a sine-like wave as Mercury orbits around the Sun. This knowledge will help geologist and space scientist to determine if Mercury has a molten core [4]. Within the framework of the MORE rotation experiment, the estimation of Mercury obliquity and libration amplitude can be accomplished by collecting pictures of Mercury's surface by using the HRIC camera. The Mercury libration and rotation axis can then be measured by analyzing image pairs relative to the same region and taken at different epochs by identifying features in the two images [3]. However, several constraints must be considered for a successful application of the method. For example, the probe will approach Mercury in a polar orbit, which will make the pictures of its surface taken at different altitude, thereby with different scale factors at different resolutions. Moreover, the estimation of the libration amplitude needs to be accomplished at epochs where the amplitude is high enough to be appreciated by the instruments. Here, the difference in phase angle (the angle between the Sun, the point on the surface and the observer) will bring images affected by different illumination conditions. Until now, only NASA Mariner 10 in the mid 70s and now Messenger have been able to acquire close up images of the Mercury surface. The former probe has provided images with a limited resolution of about 45% coverage of Mercury during the three flybys. The latter will enter into a polar orbit around Mercury soon, giving much more details on the Mercury surface.

In order to study and test the most effective pattern matching algorithms that meet the required accuracy in estimating the libration, a dataset of image pairs sharing the same portion of the planet's surface, taken at different epochs during the orbit, must be considered. The lack of both high resolution and different illumination conditions in the currently available images of the Mercury surface, requires to have at least synthetic images at one's disposal. In this work, we present the simulation environment used to generate synthetically all the possible image pairs. Starting from the generation of an earlier approximation of the Mercury surface, more morphological features (mainly craters) are added to the surface to generate the final Digital Elevation Model (DEM). The synthetic image is then raytraced according to the orbital parameters of the imaging system.

2 Previous Work

In [5], the authors simulated the pattern matching of albedo features and craters for a Mercury orbiter using synthetically generated images. Their study concluded that sub-pixel accuracy could be achieved, if images would be obtained at phase angles greater than 5 deg and less than 55 deg for albedo spots and if the difference between the phase angles are smaller than 35 deg for craters. The albedo spot features they employed are a resampled version of an albedo map of Deimos, the smaller and outermost of Mars' two natural satellites. Nevertheless, there is no evidence that albedo spots at such a small scale can be found on Mercury. Also, the craters modeled and rendered by authors are used together with a trivial and non automatic pattern matching technique. Also, the restriction imposed by the orbital constraints and Sun phase angles limits the number of image pairs available for the rotation experiment. Authors assume that the overall accuracy required for the libration and rotation estimations can be met, if at least 25 image targets are observed repeatedly over the nominal mission of 360 days.

In [6], the authors simulated the surface of Mercury in order to assess the accuracy of the pattern matching technique. The authors make use of DEMs taken from other planets and a synthetically generated DEM, subsequently raytraced with POV-Ray [8] under different illumination conditions. Results show that a match with sub-pixel accuracy can be attained on 95% of the trials on real DEMs and only in 87% of the cases with the synthetic DEM. Also, as authors state, the synthetic DEM suffers from being not sufficiently realistic. The matching of pairs is possible if the Sun *elevation* (the angle between the Sun position and the surface horizon plane) is not at the zenith, it is greater than 10 deg and the Sun *azimuth* (the angle between North and the projection of the Sun vector in the horizon reference plane) between the images is less than 30 deg.

As far as the surface features are concerned, the authors in [5], [2], [3] have chosen the albedo features as the best candidate for the pattern matching stage. When present, these features are due to changes in the reflectance of the surface material, the opposite of the variations in relief that characterize all the morphological surface features. However, the work in [12] points out that even if small scale albedo spot could be present on the Mercury surface, and even though a global map of albedo spot can be built by examinations of the currently available Mercury images, the size of such features is still prohibitive for the small field of view of the BepiColombo narrow angle camera.

3 The Method

3.1 The Surface Features Considered

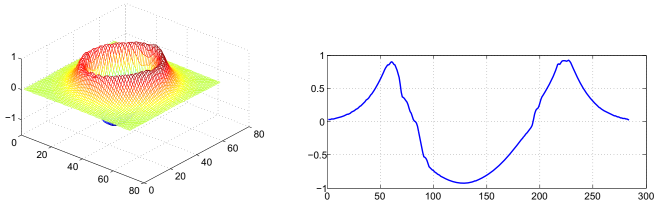
The albedo features have raised importance due to their photometric appearance being not related to geomorphological structure of the terrain. For a long time these have been desirable features for visual inspection since they allow the observer to identify the place on the planet being watched. Erroneously, their independence from any morphological structure has been considered useful for automatic pointing, or even for automatic tracking. As a matter of fact, the assumption that the albedo features are reliable features that can be employed for accurate pattern matching has to be reconsidered. First, classic albedo features of Mercury are planetary wide, they come from telescope observations and are motivated by visual observation. Accordingly, they are unfeasible to be tracked by a small FOV camera like the one used in the BepiColombo mission, designed to capture surface details. Perhaps, other local, "small scale", albedo features might be present on the Mercury surface although neither their existence nor their location is documented. Second, due to the presence of morphological features like craters and faults, the albedo may undergo shadowing effects as the Sun position changes, thus their appearance changes photometrically, being not invariant for our purposes.

On the other hand, craters are widely spread over the surface of Mercury, similarly to the Moon, although the higher gravitational field of Mercury and its closeness to the Sun implies a higher impact velocity of bodies reaching the surface. The distribution of crater sizes on Mercury extends to a wide range of diameters, from very high impact basins of the order of thousands of kilometers to very small craters of few meters [9]. According to the change in craters' appearance, due to different illumination conditions, a suite of pattern matching algorithms can be built to target each class of differences in illumination. Moreover, as long as these changes keep limited, the pattern matching can be accomplished by using local features tracked between the image pairs, either using the Lukas Kanade feature Tracker (KLT) [14] or more structured feature descriptors such as Scale Invariant Feature Transform (SIFT) [13] which take into account also the difference in scale. Therefore, our primary choice is to target all the surface features present in the images according to each class of changes in illumination and scale. In this work we target craters as the main image features, albeit not directly as *pattern matching features.*

3.2 Image Generation

The simulation system builds a DEM of the Mercury surface by simulating a set of impact craters on an initial terrain model. Craters are randomly distributed on the surface according to a set of parameters defining their diameters and the density distributions. Each crater is modeled by combining impact crater models [11,9,10] and surface fractal details.

Mercury presents mainly two types of craters: simple craters, bowl-shaped structures formed by the smallest impact bodies which create sharp rims, and complex craters, characterized by terraced walls due to subsurface faults which present a central peak of material brought up from beneath the surface. The transition diameter between simple and complex crater is 10.3 km according to [11]. The craters are modeled by using Gaussian shapes. A number of slightly shifted 2D Gaussian shapes have been superimposed in order to obtain an irregular border on top of the crater (see Fig. 1). These shapes have been cut starting from a 17% base level added by a random component. The inner side of the crater has been modeled by reversing each Gaussian shape, after doubling its sigma, in order to simulate a smooth impact area. Complex craters have been also modeled by superimposing a central peak in the impact area. Fig. 1 shows shape and profile of a simple crater generated by this method. The synthetic image generation procedure is outlined as follows:



(a) (b)

Fig. 1. Simple crater model (a) and its profile (b)

1. a fractional Brownian motion (fBm) algorithm generates the plain surface according to a $1/f^n$ pink noise, where f is the frequency and n the argument. This procedure permits to define the coarser texture level free from any other feature (i.e., craters, bumps, scratch, etc...) at each scale;
2. a random number of craters is added to the surface, with variable diameter size according to the parameters given;
3. step 1 and 2 are repeated by a prefixed number of times for each scale, until the resulting DEM is generated.

The first step allows to generate the main coarser irregular surface, and at each iteration it smooths the previous generated DEM in order to simulate different ages of impacts on the surface. This behavior is carried out by increasing the argument of the pink noise, at each scale.

A number of craters is then added randomly to the coarse DEM shape. In particular, the density of craters is within three predefined ranges, according to three different "main diameters" parameters. The latter have been estimated according to [9,10] and by analyzing the size of craters of images acquired by Messenger and Mariner10 space missions, also taking into account the ground resolution of the BepiColombo HRIC camera. Each "main diameter" pertaining to each generation stage has its own variation parameter which adds a random crater diameter within each range. It is worth noticing that, at the time being, no data is available in order to estimate what kind of features would be visible by the HRIC camera sensor, especially at the periherm side. In fact, considering that for a HRIC image the ground size is between about 10 km and 40 km at full resolution, no images at such resolution have been acquired by space probes so far. Nevertheless, small impact craters and terrain features, which can be properly modeled by fBm, are likely to be expected. The proposed simulator is able to achieve a balance in the output scene when considering these two components.

The synthetic DEM of Mercury generated by our algorithm is raytraced by a public domain rendering program [8]. At present, the camera is modeled as a pinhole camera, since optical distortions and aberrations will be already corrected in the input images. Generated images are further convolved with a nominal Point Spread Function (PSF) which has been assumed to be Gaussian with a FWHM of 1.1 pixels. This will be effectively modeled after the real camera PSF estimation. The Field Of View (FOV) of the camera is 25 mrad, which gives projections of the planet's surface having a ground size between 10 to 40 km according to the altitude of the probe. The sensor size is of 2048 × 2048 pixels at full resolution. The satellite position is expressed in latitude and longitude of the nadir point O (the point on the surface along the vertical direction according to the gravity field), together with the altitude relative to the surface ground. The Sun illumination direction is modeled by two angles: elevation and azimuth.

4 Results

4.1 Synthetic Image Generation

In Fig. 2, a rendered DEM is presented showing the generation of images with different illumination angles as well as changes in scale (last row). The first row show very a high change in both elevation angle (67 and 20 deg) and azimuth (141 and 72 deg). The shadows apper more emphasized on the second image. On the second row, Sun elevation varies slowly (ten degrees) while azimuth angle changes prominently from 12 deg (c) to 224 deg (d). Here we can see how the appearance of craters have been slightly modified. In fact, the projected shadows on the surface change the shape of the crater rims. In the third row, two images of the same DEM have been acquired with varying altitude, (e) being taken at 700 km and (f) at the nominal distance of the periherm side (400 km) with a difference of 100 deg in azimuth angles and by keeping constant the elevation angles. From these images it is readily evident how the azimuth angle plays a relevant role in affecting the subsequent pattern matching stage. Craters' shadows are altered both in the inner and outer regions. This behavior, while being almost impractical to be tackled by correlation-based and feature descriptors matching methods, could be faced by shape from shading and combined matching algorithms.

In Fig. 3, the terrain generated is heavily cratered. Also the illumination angle changes slightly from frame to frame. In fact, elevation is near horizon for

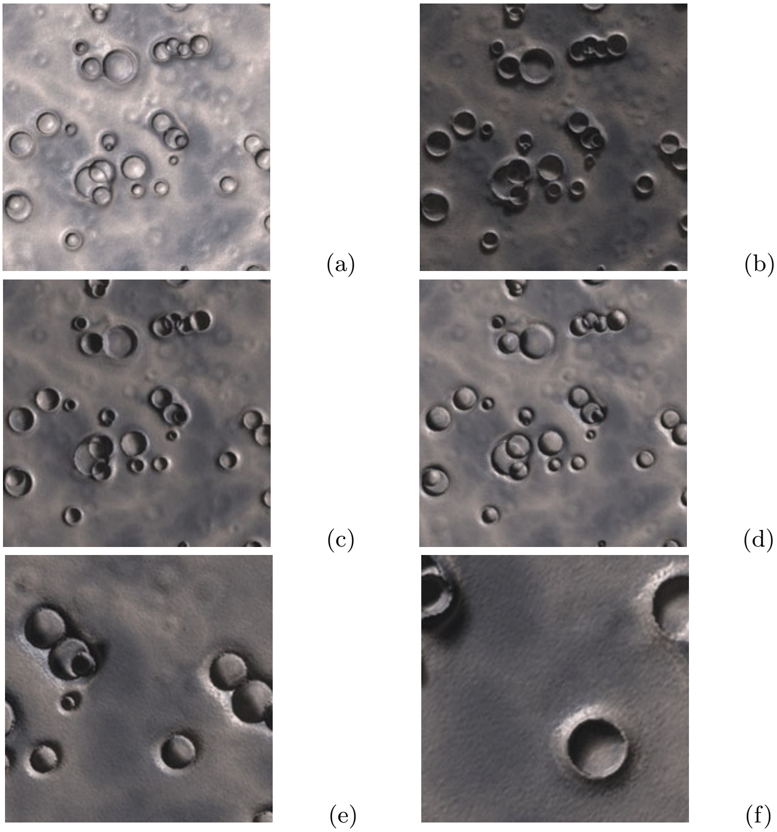


Fig. 2. Simulated images of Mercury surface at different illumination angles

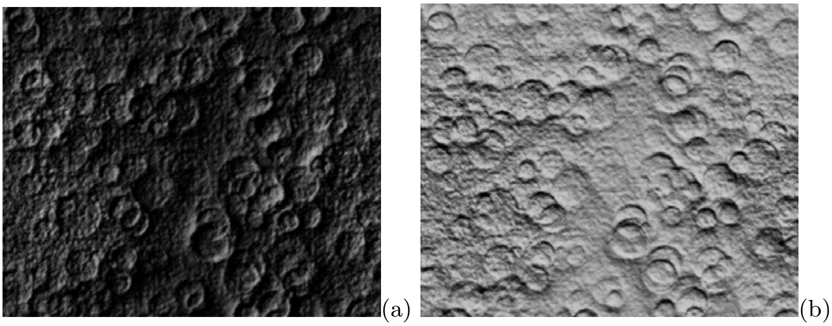


Fig. 3. Simulated images of Mercury surface in a heavily cratered area

the first image and about 42 deg for the second one, whilst azimuth angles are almost in opposition. Here, the shadows projected by craters and surface's erosion make practically every pixel of the two frames unrelated. Thus, also in this example, correlation-based as well as shape-based pattern matching algorithms would likely fail.

4.2 Feature Tracking

Here, a preliminary study on the extraction and matching of features is reported. In Fig. 4(a), the SIFT features extracted and matched from Figs. 2(d) and 2(e) are shown. The two frames show a change in altitude of about 800 km, a change in illumination angles of 10 deg for elevation and no change in illumination

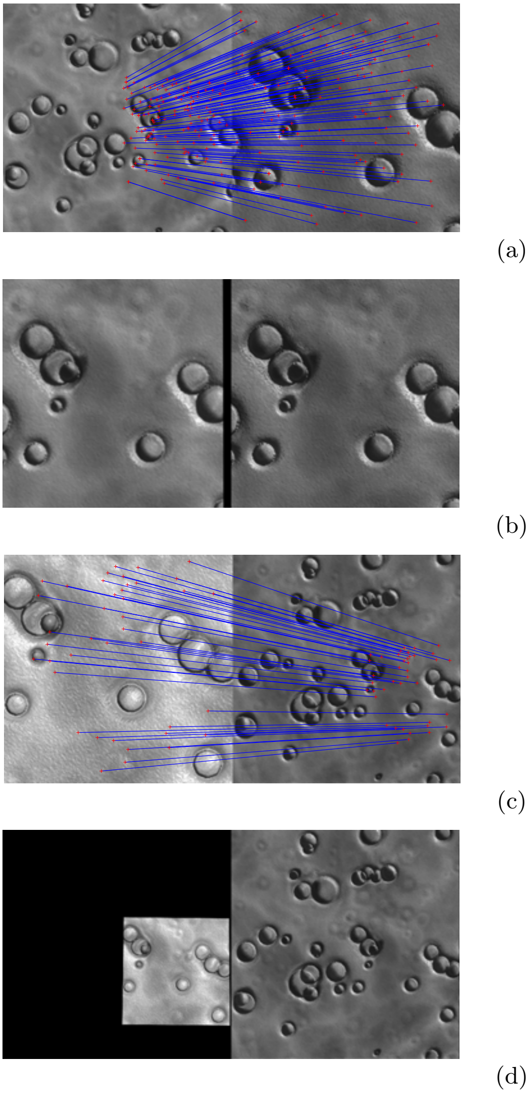


Fig. 4. Matched SIFT features between two frame couples (a,c) taken at different scale and illumination angles: difference in elevation is 10 deg in (a) and more than 60 deg in (c). In (b,d) the first frame is reprojected according to the estimated H (see text).

azimuth. Since the camera moves according to the known orbital parameters and from that altitude an approximation of a planar surface with respect to the small FOV could be made, the two views are related by an homography transformation, which maps points in the first frame to the corresponding points on the second frame. In a more realistic case, also the dispalcement of the camera centers will be included into the transformation model, thereby relaxing the panar surface assumption. The homography H recovered by using RANSAC [15] is used to estimate the scale and the position of the corresponding pixel in the two images. The first image reprojected on the domain of the second is shown in Fig. 4(b). The estimation accuracy is quite good, being the imposed offset (133 and 59 pixel in the horizontal and vertical directions, respectively) effectively recovered with sub-pixel accuracy, even when elevation angles vary of more than 60 deg (Fig. 4(c,d)). However, as expected, trials conducted with very low or very high elevations and with changes in the azimuth angle show that this feature matching fails. These are among the numerous cases that have to be tackled with an ensemble of matching algorithms built on purpose.

5 Conclusion

The BepiColombo rotation experiment, which aims at estimating the obliquity and libration of Mercury, is a crucial experiment in order to assess many important implications on our solar system. The recovery of the libration amplitude can be achieved by matching surface features extracted from image pairs taken at different epochs and thereby affected by illumination and scale artifacts. The lack of detailed image pairs available from other space missions to Mercury limits the possibility to assess the accuracy of the methods we propose. In this work a simulation framework aiming to both generate realistic images of the Mercury surface as well as devising pattern matching algorithms is presented. The framework is able to generate synthetic images according to the orbital parameters of the space mission, thereby allowing to test the pattern matching algorithms under different scale and illumination conditions, as close as possible to the real case scenario. Finally, some preliminary matching results dealing with surface features extracted from our synthetic images taken at different altitude and affected by illumination changes show that sub-pixel accuracy is possible even for high changes in Sun elevation angle. Also, as expected, trials conducted with very low or very high elevations and with changes in the azimuth angle have to be tackled with an ensemble of matching algorithms built on purpose.

References

1. Schulz, R., Benkhoff, J.: BepiColombo: Payload and mission updates. Advances in Space Research 38(4), 572–577 (2006)
2. Iess, L., Boscagli, G.: Advanced radio science instrumentation for the mission BepiColombo to Mercury. Planet. Space Sci. 49, 1567–1608 (2001)

3. Iess, L., Asmar, S., Tortora, P.: MORE: An advanced tracking experiment for the exploration of Mercury with the mission BepiColombo. Acta Astronautica 65, 666–675 (2009)
4. Jehn, R., Corral, C., Giampieri, G.: Estimating Mercury's 88-day libration amplitude from orbit. Planetary and Space Science 52(8), 727–732 (2004)
5. Jorda, L., Thomas, N.: The accuracy of pattern matching techniques for the radio science experiment of ESA's Mercury Cornerstone mission. Preliminary study at Max-Planck Institute for Aeronomie, Katlenburg-Lindau, Germany (2000) (unpublished)
6. Vacanti, G., Buis, E.J., Beijersbergen, M.: BepiColombo: Study on Techniques and Accuracies of Pattern Matching of Remote Sensing Images. Tech. rep., Cosine Research ESA (2005)
7. Peale, S.J., Yseboodt, M., Margot, J.: Long-period forcing of Mercurys libration in longitude. Icarus 187, 365–373 (2007)
8. Persistence of Vision Raytracer (2004), http://www.povray.org/
9. Gault, D.E., Guest, J.E., Murray, J.B., Dzurisin, D., Malin, M.C.: Some Comparisons of Impact Craters on Mercury and the Moon. J. Geophys. Res. 80(17), 2444–2460 (1975)
10. Melosh, H.J., Ivanov, B.A.: Impact Crater Collapse. Earth Planet. Sci. 27, 385–415 (1999)
11. Pike, R.J.: In: Vilas, F., Chapman, C.R., Matthews, M.S. (eds.) Mercury, pp. 165–273. University of Arizona Press, Tucson (1988)
12. Denevi, B.W., Robinson, M.S.: Mercury's albedo from Mariner 10: Implications for the presence of ferrous iron. Icarus 197(1), 239–246 (2008)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
14. Tomasi, C., Kanade, T.: Detection and Tracking of Point Features. Carnegie Mellon University Technical Report CMU-CS-91-132 (1991)
15. Fischler, M.A., Bolles, R.C.: Random sample and consensus: A paradigm for model fitting with application to image analysis and automated cartography. Comm. of the ACM 24, 381–395 (1981)

A Novel T-CAD Framework to Support Medical Image Analysis and Reconstruction

Danilo Avola[1], Luigi Cinque[1], and Marco Di Girolamo[2]

[1] Sapienza University of Rome, Department of Computer Science
Via Salaria 113, 00198 Rome, Italy
{avola,cinque}@di.uniroma1.it
http://w3.uniroma1.it/dipinfo/english/index.asp
[2] Sapienza University of Rome, S.Andrea Hospital, Department of Diagnostic Science
Via di Grottarossa 1035-1039, 00189 Rome, Italy
digirolamomarco@hotmail.com
http://www.ospedalesantandrea.it/index.php

Abstract. The current medical imaging devices allow to obtain high resolution digital images with a complex informative content expressed by the textural skin that covers organs and tissues (hereinafter objects). These textural information can be exploited to develop a descriptive mathematical model of the objects by which to support heterogeneous activities within the medical field.

This paper describes our developed framework based on the texture analysis by which to mathematically model every object contained in the layout of the total body NMR images. By every specific model, the framework automatically also defines a connected application which supports, on the related object, different fixed targets, such as: segmentation, mass detection, reconstruction, and so on.

Keywords: Framework, CAD, medical image, texture analysis, pattern recognition, feature extraction, segmentation, classification.

1 Introduction

The current medical imaging devices (e.g. Nuclear Magnetic Resonance (NMR), Positron Emission Tomography (PET)) allow to obtain digital images with high level details. These images have a complex informative content that goes beyond the simple visual representation. More specifically, by observing the relationships between clusters of pixels (i.e. the texture) of the skin that covers the objects can be brought out the meaningful features by which to describe the morphological structures related to objects themselves. These features (i.e. textural information) are exploited to develop a mathematical model, of the objects, able to support different activities within the medical field. In the last years, there have been many efforts to conceive intelligent and automated systems to support several critical medical tasks (e.g. analysis, masses identification). The making of the current Decision Support Systems (DSSs), better known in this area as Computer Aided Diagnosis (CAD) systems, are still not very effective tools.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 414–423, 2011.

This paper describes a generalized texture analysis and reconstruction approach, based on our previous experiences shown in [1], [2] and [3], by which to support the total body NMR images. In particular, the paper details our developed ***T****exture based* ***C****omputer* ***A****ided* ***D****iagnosis* ***Framework*** (T-CAD Framework) which, on one side, supports the mathematical modeling process of each organ and tissue represented inside the total body NMR image layout and, on the other hand, automatically defines, by every specific model, an ad-hoc application by which to support a set of fixed targets useful to aid the medical specialist in a specific context (e.g. craniopharyngioma identification on NMR encephalic images). Observe that every mathematical model (e.g. craniopharyngioma model) achieved during the modeling process is defined one and for all.

There is an extensive literature focused on the different aspects of the medical image processing. These approaches are based on several principles related to the *image understanding*, but there are no works able to exploit the morphological structure (given by the texture analysis) of the objects with the aim to provide both their descriptive mathematical model and related dedicated T-CAD system. The effectiveness of the proposed approach is shown both experimental results and the ad-hoc model-driven applications.

A first work that has driven some our choices in image analysis approach is shown in [4], where the authors describe an automatic segmentation framework, for brain Magnetic Resonance Imaging (MRI), based on the combination of atlas registration, fuzzy connectedness segmentation, and parametric bias field correction. Another work that has supported some aspects related to the textural image filters is detailed in [5], where the authors present a novel co-occurrence matrixes based approach able to discriminate the texture belonging to different kind of images by considering the statistical representation of the structural texture primitives (i.e. textons). Another useful approach is presented in [6], where the authors presents an approach to automate the myocardial contours detection in order to optimize the detection and the tracking of the grid of tags within myocardium. A further work that has supported our framework is shown in [7]. In this work the authors present a real mixed statistical model based on a region-driven curve evolution algorithm. An original approach that has assisted some aspects of our texture based image processing is described in [8], where the authors introduce a novel algorithm to achieve automatic texture based segmentation of organs in MRIs of the abdomen. An innovative approach, which has inspired some solutions of our textural model, is detailed in [9], where the authors present a robust multi-resolution statistical shape model algorithm. A last remarkable work that has aided our collaboration process between textural filters is presented in [10], where the authors explore the use of the co-occurrence matrixes to extract textural features from medical images.

The paper is organized as follows. Section 2 illustrates the main architectural aspects of the T-CAD Framework. Section 3 presents a concrete case study including the related mathematical model. Section 4 introduces and discusses the main experimental results. Section 5 concludes and plans the future work.

2 The T-CAD Framework Architecture

The developed T-CAD Framework is a smart tool that allows skilled user to define a texture based ***Mathematical Model*** (MM) of every organ and tissue represented inside the NMR total body image layout. A specific MM (e.g. of the brain) is simply the set of formalized mathematical classes which represent the different ***basic objects*** contained in the related NMR image (e.g. cerebral tissue, abnormal mass, background). For every MM the framework supports the building of a dedicated T-CAD system to support a specific medical image analysis process (e.g. NMR encephalic image segmentation and mass recognition).

The next two sub-sections respectively show the general approach, and the main textural filters used within the framework.

2.1 The Region Based Algorithm

The *Region Based Algorithm* (RBA) can be considered the core of the whole T-CAD Framework. Actually, its main aspect regards the building of all the mathematical classes that made up a specific MM, inasmuch the dedicated system generation is only a technical application of the related MM on a dataset of source images. For this reason, the explanation will mainly detailed the definition of the MM, and finally will be highlighted the system generation process. Figure 1 shows the simplified architecture of the region based algorithm.

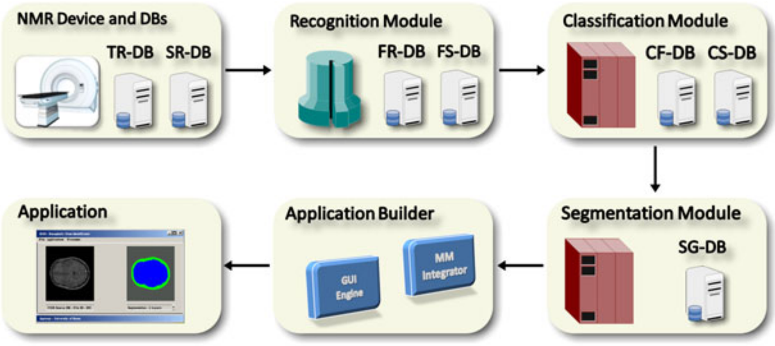


Fig. 1. Region Based Algorithm Architecture

The *first panel* (NMR Device and DBs) highlights that our framework works on two kind of DB. The first (**TR**aining-DB) is used when the skilled user has to build a new MM. For this reason, the populating of the TR-DB follows a rigid protocol which has to ensure different qualitative and quantitative requirements relatively to the informative content of the related images. The second DB (**S**ou**R**ce-DB) points out that, once obtained the related MM, it is possible to analyze every kind of source image coming directly form the NMR device.

The *second panel* (Recognition Module) highlights the ***recognition process*** on every image belonging to the TR-DB, where an adaptable elaboration window runs across the image to perform a feature extraction process based on a suitable set of textural filters (i.e. *features vector*). The start step, supported by two specific filters: *Entropy* and *Homogeneity* (see next sub-section), is to fix the window size both maximizing the number of image zones with high entropy levels and minimizing the number of neighboring heterogeneous zones. The final step of the process is to exploit the found fixed window to analyze, by the features vector (which filters are stored in the **F**ilte**R**-DB), every image area (in top-bottom and left-right way). For each image the analysis process produces a feature map, the set of feature maps defines a feature space (stored in the **F**eature **S**pace-DB), all of which, suitably interpreted, provide the mathematical class of every chosen basic object. In fact, by studying the correlation related to all the feature spaces of each basic object can be defined a preliminary mathematical model of everyone. Afterwards, the exhaustive comparison of each different preliminary model allows skilled user to find the textural relationships to univocally describe each basic object. This description is the MM, which can be considered as the set of the mathematical classes that define itself. Observe that the MM definition is a supervised process in which the final formalization of the numerical classes is still left to the human specialist. Moreover, the process only occurs when a skilled user wants to build a new MM.

The *third panel* (Classification Module) points out the ***classification process*** which exploits the MM found in the previous module (and stored in the **C**lass**F**ormat-DB) with the aim to analyze the source images coming form the NMR device (SR-DB). This process follows the same approach shown in the recognition module, but its purpose is to classify any zone of every source image according to a selected MM. The module works following two different steps. During the first any image zone is analyzed (by the just mentioned elaboration window) to classify it according to one of the formalized mathematical classes belonging to the related MM. Afterwards, homogeneous image zones are suitably marked and merged. During the second step the module assigns to one of the formalized classes, by a statistical distribution algorithm, the possible image zones that have been not classified at all.

The *fourth panel* (Segmentation Module) highlights the ***segmentation process*** where every classified source image is properly labeled in order to to provide an immediate visual impact to the user. Every segmented image is arranged in a suitable DB (**S**e**G**mented-DB).

The *fifth panel* (Application Builder) points out the ad-hoc application builder. In particular, the *MM Integrator* will include the definition of the selected MM inside the application. Observe that the framework allows user to include more than one specific MM. The *GUI Engine* highlights that the main mechanisms related to the data presentation (e.g. visualization engine, interaction properties) are the same independently from the specific application.

The *sixth panel* (Application) shows an example of a created application which performs the encaphalic NMR image segmentation.

2.2 The Textural Image Filters

The image filters adopted to support the RBA have been suitably chosen and/or created to define the basic textural informative content of the tissues and organs belonging to the human body. Actually, our strong belief is to have found a general approach adoptable for every object represented by the NMR images, which, at the moment, has been refined to define the textural morphological structures of four specific objects: brain, heart, liver and bony structure.

Our approach subdivide the image filters within three different graphical classes, each one able to characterize a specific informative layer of the mentioned objects: *informative class*, *texture class* and *pattern class*.

The *informative class* is made up by those *first order statistic* image filters which distinguish between zone with and without relevant information content. In our experience, the following two set of filters represent the main best suitable ones: *N-Order Moment* (M_{n_1}) and *N-Order Central Moment* (C_{n_2}):

$$M_{n_1} = \sum_{i=0}^{N} i^{n_1} \times p(i), C_{n_2} = \sum_{i=0}^{N} \left(i - M_{n_1}\right)^{n_2} \times p\left(i\right) \tag{1}$$

Where: *p(i)* represents the probability that the gray level value i appears inside the elaboration window. The following constraints must be respected:

$$0 \leq p\left(i\right) \leq 1 \; \forall i \in [0..255] \subset \mathbb{N}, \sum_{i=0}^{N} p(i) = 1, n_1, n_2 \in \mathbb{N}, N = 255 \tag{2}$$

The M_{n_1} and C_{n_2} set of filters respectively measure, on different textural graphical layers, the consistent quantity and the semantic readability of the information related to different image zones.

The *texture class* is made up by those *second order statistic* image filters which measure the macro and micro textural structures. Our empirical studies have allowed to detect the following four set of filters: *Homogeneity* ($Hg(d)$), *Contrast* ($Ct(d)$), *Inverse Difference* ($Id(d)$) and *Entropy* ($En(d)$):

$$\begin{aligned} Hg(d) &= \sum_{i=0}^{N}\sum_{j=0}^{N} \left[p_d(i,j)\right]^2, Ct(d) = \sum_{i=0}^{N}\sum_{j=0}^{N} \left|i,j\right| \left[p_d(i,j)\right]^l \\ Id(d) &= \sum_{i=0}^{N}\sum_{j=0}^{N} \frac{p_d(i,j)^m}{1+(i-j)^k}, En(d) = -\sum_{i=0}^{N}\sum_{j=0}^{N} p_d(i,j)\log_n(p_d(i,j)) \end{aligned} \tag{3}$$

Where: $p_d(i,j)$ represents the probability that two points with distance d have respectively i and j gray value. The following constraints must be respected:

$$\begin{aligned} &0 \leq p_d(i,j) \leq 1, \forall (i,j) \in [0..255] \times [0..255] \subset \mathbb{N}^2 \\ &\sum_{i=0}^{N}\sum_{j=0}^{N} p_d\left(i,j\right) = 1 \, , d \in [1..8] \subset \mathbb{N}, l, m, k, n \subset \mathbb{N}, N = 255 \end{aligned} \tag{4}$$

The $Hg(d)$ set of filters measure the degree of uniformity of the different image zones, where high or low values within the feature maps respectively highlight light or wide changes of the textural structures. The $Ct(d)$ set of filters express how roughly occur the mentioned structural changes, where high values, of the related feature maps, point out fast continuous changes within the image zones, on the contrary, they give low values. The $Id(d)$ set of filters provide the measure of the transition between different basic objects, where low values typically highlight a boundary zone. The $En(d)$ set of filters is used to detect the randomness level within the considered image zone, where the complex changes in the random distribution of the grey levels are directly proportional to the given values.

The *pattern class* is made up by those *second order statistic* image filters which measure the pattern structures. Our experimental observations have supported to identify the following two set of filters: *Correlation* ($Cr(d)$), *Difference Entropy* ($De(d)$):

$$Cr(d) = \sum_{i=0}^{N}\sum_{j=0}^{N} \frac{(i-\mu_x)(j-\mu_y)p_d(i,j)^l}{[\sigma_x\sigma_y]^m}, De(d) = -\sum_{i=0}^{N} p_{x-y}(i)\log_n[p_{x-y}(i)] \qquad (5)$$

Where:

$$\begin{aligned}
\mu_x &= \sum_{i=0}^{N}\sum_{j=0}^{N} i\left(p_d(i,j)\right), \sigma_x = \sqrt{\sum_{i=0}^{N}\sum_{j=0}^{N}(i-\mu_x)^2\left(p_d(i,j)\right)} \\
\mu_y &= \sum_{i=0}^{N}\sum_{j=0}^{N} j\left(p_d(i,j)\right), \sigma_y = \sqrt{\sum_{i=0}^{N}\sum_{j=0}^{N}(i-\mu_y)^2\left(p_d(i,j)\right)} \\
p_{x-y}(k) &= \sum_{i=0}^{N}\sum_{j=0}^{N}[p_d(i,j)]^q, where|i-j| = k
\end{aligned} \qquad (6)$$

The $Cr(d)$ set of filters are usually used to recognize definite patterns within texture zones previously identified, while the $De(d)$ set of filters is adopted to detect the different components (i.e. parts of a same pattern) of different basic objects.

Every class of image filters is based on our variation of the current co-occurrence matrix concept, where the elaboration process considers all the search directions. In particular, every couple of pixels, able to increase a coefficient of the matrix, is chosen considering a pixel (i.e. center of a circle) within the elaboration window and the connected pixel which is at the boundary of the circumference related to the fixed radius (i.e. the d parameter).

All the parameters belonging to the set of filters are customized, within the recognition process, according to the specific contextual medical domain and related tasks. Besides, the shown filters are often used on more than one level of the Gaussian Pyramid ([11]) by which to enrich the resolution of the texture described through the MMs.

3 Case Study: Brain

This section shows a concrete case study in which a suitable mathematical model has been defined according to specific targets, and where a model-based application has been created to support the related image analysis process. Actually, the framework essentially produces the same application in which only the mathematical model is replaced every time.

Observe that the morphological structures of organs and tissues are very different, each one of they can be better emphasized according to a specific kind of NMR image (e.g. T_1, T_2, *proton density*). For this reason, a brief specification of the DICOM (*Digital Imaging and Communication in Medicine*) image format is given in relation to a specific case study.

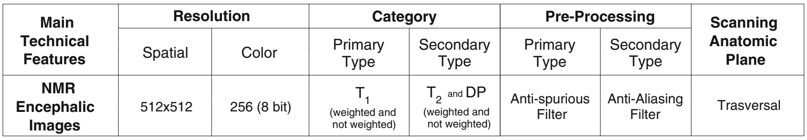
3.1 NMR Encephalic Mass Identificator

The ***E**ncephalic **M**ass **I**dentification* (EMI) application has been created with the aim to aid medical specialists during the mass identification within the encaphalic NMR images. In particular, the following targets were established:

a. *segmentation* of the image layout in three basic objects: cerebral tissue, rest of the image (i.e. muscular and bony structure) and background;
b. *identification* within the cerebral tissue of the abnormal masses (e.g. gliomas, craniopharyngiomas, medulloblastomas);
c. *classification* of the found abnormal masses as craniopharyngioma pathology distinguishing it from other kinds of primary cerebral tumors.

A careful analysis of the set of pixel that composes the encephalic NMR images has detected in the transversal T_1 weighted the more suitable ones to better highlight the textural morphological structures of the objects, according to the fixed targets. Table 1 shows the main technical features of the images.

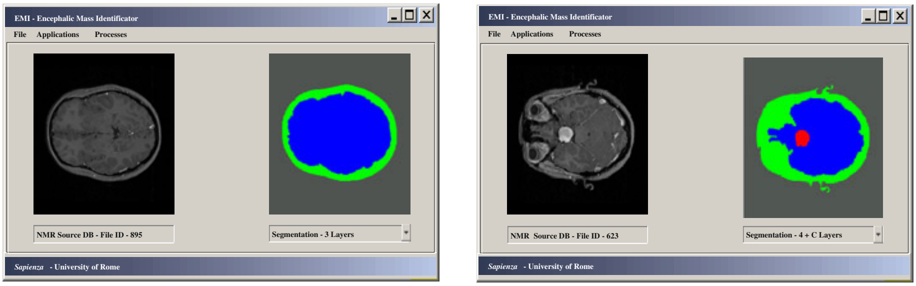
Table 1. NMR Encephalic Images: Main Technical Features



Main Technical Features	Resolution		Category		Pre-Processing		Scanning Anatomic Plane
	Spatial	Color	Primary Type	Secondary Type	Primary Type	Secondary Type	
NMR Encephalic Images	512x512	256 (8 bit)	T_1 (weighted and not weighted)	T_2 and DP (weighted and not weighted)	Anti-spurious Filter	Anti-Aliasing Filter	Trasversal

The Table 1 highlights that the images belonging to the others categories have been used to refine and optimize the textural feature extraction methodology. Moreover, it shows that two different kinds of pre-processing filters have been applied on the related source image with the aim to normalize the original gray levels. The mentioned filters do not alter the quality of the original source image, but they are only used to improve the few image zones affected from lack of information (i.e. localized noise).

The two screenshots shown in Figure 2 point out the segmentation process of the EMI application on two encaphalic NMR images. In particular, the first screenshot (left) shows an image in which the three layers related to basic objects are found (a). The second one (right) highlights the recognition of four layers


EMI - Encephalic Mass Identificator
File
Applications
Processes
Sapienza - University of Rome
EMI - Encephalic Mass Identificator
File
Applications
Processes
Sapienza - University of Rome

NMR Source DB - File ID - 895

Segmentation - 3 Layers

NMR Source DB - File ID - 623

Segmentation - 4 + C Layers

Fig. 2. EMI Application: NMR encephalic segmentation

where it is also identified an abnormal mass classified as craniopharyngioma pathology (a, b and c).

The image analysis process of the EMI application has been based on the following ***C**ranio**ph**aryngioma **M**athematical **M**odel* (CPH-MM):

$$\begin{aligned}
&\{CPH-MM\} = (((25 \leq Cr(1) \leq 145) \vee (198 \leq Cr(1) \leq 255))\wedge \\
&((33 \leq Cr(2) \leq 180) \vee (220 \leq Cr(2) \leq 250)) \wedge ((23 \leq Cr(2) \leq 109)\vee \\
&(148 \leq Cr(3) \leq 213))) \wedge (((70 \leq De(1) \leq 112) \vee (160 \leq De(1) \leq 200))\wedge \\
&((34 \leq De(2) \leq 57) \vee (100 \leq De(2) \leq 127)) \wedge ((110 \leq De(3) \leq 200)\vee \\
&(220 \leq De(3) \leq 255)))
\end{aligned} \tag{7}$$

Actually, the EMI application needs to take into account also the ***A**bnormal **M**ass **M**athematical **M**odel* (AB-MM) which primary support the general abnormal mass detection (see [2]). Subsequently, the application, only on the abnormal mass, can apply the CPH-MM with the aim to recognize it as belonging to the craniopharyngioma pathology.

3.2 Mathematical Reconstruction

Observe that the MMs, on one side, define the formalized mathematical classes by which to represent the basic objects. On the other hand, they support a first step inside the 3D reconstruction and rendering environment, in fact the provided classes have an exhaustive informative content. A skilled user can already exploit the mathematical classes on different anatomic scanning plans to inference complex information, but one of our next steps is to implement a visual 3D rendering engine.

4 Experimental Results

This section summarizes the experimental results regarding the total body NMR image analysis, related to the more advanced current case studies on the following medical domains: brain, heart, liver, bony structure. In order to explain the experimental session, the following three general tasks have been selected according to the present medical image analysis process: *task 1*: layout segmentation, *task 2*: abnormal mass or lesion detection, *task 3*: textural characterization.

The Table 2 shows the experimental session which has been subdivided in two different phases. The first, regarding the NMR image recognition, has concerned the selection of patients (455) by which to define the set of training images (1850) to build the four basic MMs. The second, regarding the NMR image classification and segmentation, has concerned the selection of patients (615) by which to obtain a set of images (2565) to test each MM on the mentioned tasks.

Table 2. Main Case Studies: Training and Source DB

Medical Domain	Patients	Training DB Images			Source DB Images			Partial Images
		Task 1	Task 2	Task 3	Task 1	Task 2	Task 3	
Brain	270 + 385	450	680	60	975	775	40	2980
Heart	75 + 100	150	80	35	170	100	25	560
Liver	60 + 80	70	100	30	135	90	20	445
Bony Structure	50 + 50	70	85	40	100	100	35	430
Total Images	455 + 615	740	945	165	1380	1065	120	4415

Actually, the best qualitative results (more than 90% of success rate) comes from the brain MM which has a large amount of training and test images. Moreover, it has been our historical first study case. Also the remaining three models have an high success rate (between the 65% and 80%), but each one needs of a more wide experimental session. In fact, the accuracy of the MM is strongly tied to the amount of training images used to refine the mathematical classes.

5 Conclusions

This paper describes the main aspects of our developed T-CAD Framework, which exploits the textural information that covers organs and tissues, represented within the NMR total body images, to perform different activities related to the medical image processing field.

The skilled user, once established both the contextual medical domain (e.g. encephalic analysis) and the specific task (e.g. craniopharyngioma recognition) can build the related MM (i.e. the set of suitable mathematical classes) able to describe the basic objects (i.e. cerebral tissue, rest of the image, abnormal mass, craniopharyngioma and background) useful to the achievement of the same task just mentioned (i.e. craniopharyngioma recognition on any compatible image dataset). In order to achieve the task, the framework allows user to create a suitable application based on the related MM. Observe that every MM is defined once and for all. At the moment, our main work is to refine the MM of the following organs and tissues: brain, heart, liver, bony structure. Another our goal is the development of a rendering engine to visualize a readable and interactive 3D objects reconstruction.

References

1. Avola, D., Cinque, L.: Encephalic NMR Image Analysis by Textural Interpretation. In: Avanzi, R.M., Keliher, L., Sica, F. (eds.) Proceedings of the 2008 ACM Symposium on Applied Computing, SAC 2008. LNCS, March 16–20, pp. 1338–1342. ACM Press, New York (2009)
2. Avola, D., Cinque, L.: Encephalic NMR Tumor Diversification by Textural Interpretation. In: Foggia, P., Sansone, C., Vento, M. (eds.) ICIAP 2009. LNCS, vol. 5716, pp. 394–403. Springer, Heidelberg (2009)
3. Avola, D., Cinque, L., Di Girolamo, M.: Texture Based Approches to Support Medical Image Analysis. Internal Technical Report in Medical Image Proessing. DSI - Sapienza University of Rome, ITR-MIP 2010, Int. Res. on Medical Imaging (2010)
4. Zhou, Y., Bai, J.: Atlas-Based Fuzzy Connectedness Segmentation and Intensity Nonuniformity Correction Applied to Brain MRI. IEEE Trans. on Biomed. Eng. Spons. by IEEE Eng. in Med. and Bio. Soc. 54(1), 122–129 (2007)
5. Li, Q., Shi, Z.: Texture Image Retrieval Using Compact Texton Co-Occurrence Matrix Descriptor. In: Proceedings of the 11th ACM International Conference on Multimedia Information Retrieval, MIR 2010, March 29-31, pp. 83–90. ACM Press, Philadelphia (2010)
6. Histace, A., Matuszewski, B., Zhang, Y.: Segmentation of Myocardial Boundaries in Tagged Cardiac MRI Using Active Contours: A Gradient-Based Approach Integrating Texture Analysis. International Journal of Biomedical Imaging, IJBI 2009, 1–8 (2009)
7. Niranjan, J., Michael, B.: Non-Parametric Mixture Model Based Evolution of Level Sets and Application to Medical Image. International Journal of Computer Vision, IJCV 2010 1(88), 52–68 (2010)
8. Wu, J., Poehlman, S., Noseworthy, M.D., Kamath, M.V.: Texture Feature Based Automated Seeded Region Growing in Abdominal MRI Segmentation. Journal in Biomed. Sc. and Eng., JBiSE 2009 2, 1–8 (2009)
9. Schmid, J., Kim, J., Magnenat-Thalmann, N.: Robust Statistical Shape Models for MRI Bone Segmentation in Presence of Small Field of View. International Journal of Medical Image Analysis, IJMIA 2011 15(1), 155–168 (2011)
10. Tesar, L., Smutek, D., Shimizu, A., Kobatake, H.: Medical Image Segmentation Using Co-Occurrence Matrix Based Texture Features Calculated on Weighted Region. In: Proceedings of the 3rd Conference on IASTED International Conference: Advances in Computer Science and Technology, ACST 2007, pp. 243–248. ACTA Press, Phuket (2007)
11. Heeger, D.J., Bergen, J.R.: Pyramid-Based Texture Analysis/Synthesis. In: Proceeding of 22nd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1995, pp. 229–238. ACM Press, NY (1995)

Fast Vision-Based Road Tunnel Detection

Massimo Bertozzi, Alberto Broggi, Gionata Boccalini, and Luca Mazzei

VisLab - Dipartimento di Ingegneria dell'Informazione
Università degli Studi di Parma, Italy
{bertozzi,broggi,mazzei}@vislab.it,
gionata.boccalini@studenti.unipr.it
http://www.vislab.it

Abstract. When a vehicle equipped with an artificial vision system enters or exits a tunnel, the camera may temporarily suffer from reduced visibility, or even get completely blind due to quick changes in enviromental illumination.

This paper presents a vision-based system that detects approaching tunnels entrances or exits. The proposed system allows other ADAS (*Advanced Driver Assistance Systems*) to act on camera parameters to effectively avoid the tunnel blindness effect. Information regarding approaching tunnel entrance can be helpful for other sensors as well and for sensor fusion systems. In terms of path planning, this system can also inform GNSS-based systems (*Global Navigation Satellite System*), which usually do not receive any signal in tunnels, and trigger dead reckoning techniques.

The proposed system is noticeably fast and therefore well fit to be used as a background process to support other ADAS applications.

Keywords: Intelligent vehicle, autonomous driving, ADAS, tunnel detection.

1 Introduction

This article presents a system for road tunnels detection in automotive scenarios. The recognition of road tunnels may bring important benefits to both automatic driving vehicles and driver assistance systems (ADAS).

Concerning autonomous driving, tunnel detection is primarily useful in two situations:

1. to adapt the camera dynamics in advance. In fact, cameras are generally suffering from abrupt changes in brightness, so the gain control system could be improved by the prediction of the next change in lightining conditions. In [5] a contrast invariant obstacle detection method is shown, which highlights the problem mentioned above.
2. To inform the path planner about the imminent lack of reception of satellite signals. These systems have major problems while the vehicle drives into tunnels [4], therefore the approach presented in this paper can fed the path planner with information regarding decisions about the use of dead reckoning techniques.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 424–433, 2011.

Concerning ADAS, the knowledge of an approaching tunnel entrance or exit location may be useful to enable certain vehicle behaviors without the driver intervention: turning lights on or off [3], or turning on or off the wipers, or interacting with other sub-systems related to driver comfort as instrument panel lighting, and more.

The developed system, based on the taxonomy presented in [1], may be placed in the category of assisted ADAS when it informs the driver, or semi automated ADAS when it delivers decisions to vehicle instrumentation.

This paper is arranged as follows: section 2 shows the algorithm architecture and the procedure to detect a tunnel entrance and exit. Results are presented in section 3 with some critical cases, and finally the conclusions will be drawn in section 4.

2 Algorithm Description

The algorithm receives as input a monocular grayscale image, then it performs a downsample procedure to reduce the computational load of the following stages.

The system can be represented as a finite state machine (FSM) with four states; for each image only one state is possible.

Figure 1 shows the finite state diagram architecture of the algorithm and its possible transitions. Each state of the FSM identifies the environment or situation met by the vehicle:

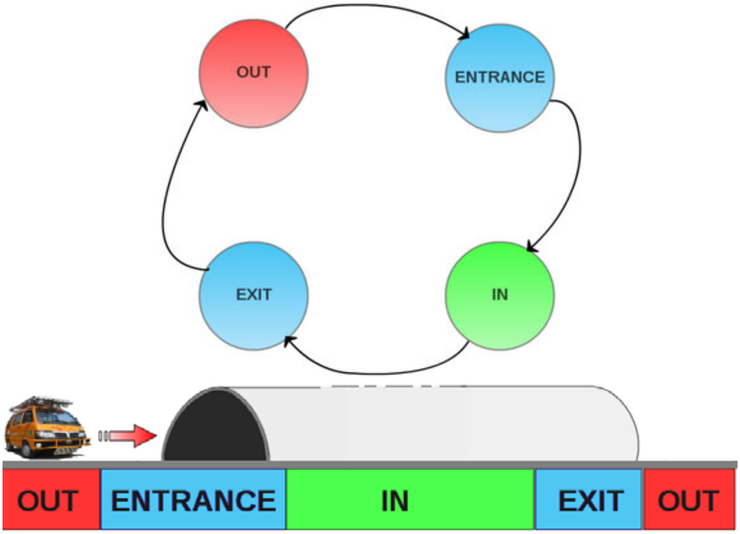


Fig. 1. Finite state machine diagram of the algorithm and representation of the various states

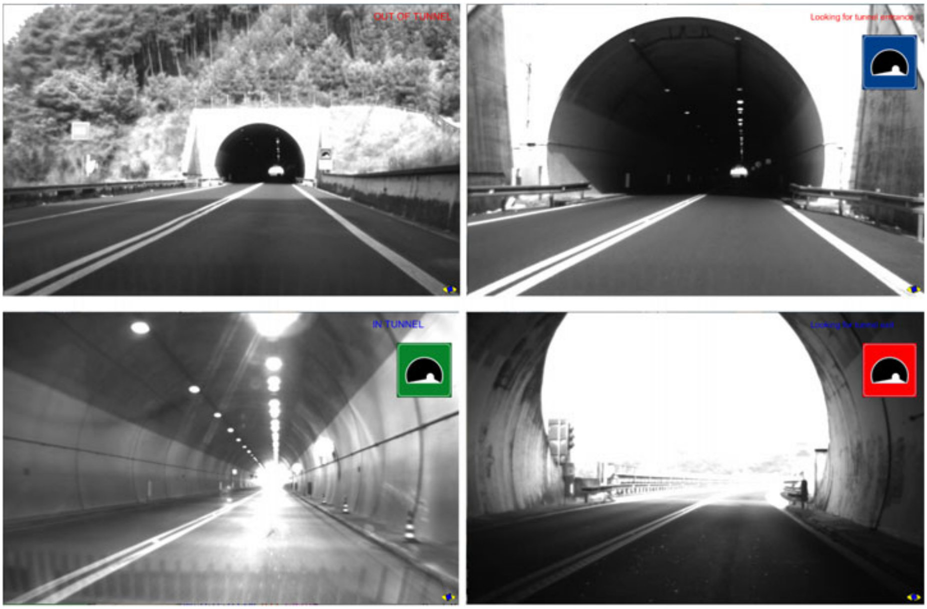


Fig. 2. Images related to FSM states with informative signals: out, entrance, in, exit

- *OUT* state: the vehicle is not inside a tunnel;
- *ENTRANCE* state: the vehicle is close to the entrance of a tunnel. In this state any procedure for entering in the tunnel can be activated, i.e. sending information about the imminent change of light conditions to the camera controller;
- *IN* state: the vehicle is inside the tunnel; the transition from outside to inside has ended.
- *EXIT* state: this is the opposite to the entrance state, from here the vehicle is approaching the tunnel exit and the system can inform the camera controller of the imminent change of light conditions.

Figure 2 schematically shows the four states and where they are activated during the passage in a tunnel.

2.1 Main Execution Flow

The main execution flow of the system mimics the FSM and, when the vehicle is travelling through a tunnel, leads the transitions amongst all states (see fig 3). Each frame is processed by a validation procedure that classifies the frame state and matches it against the current state.

Starting with the *OUT* state, the next one is *ENTRANCE*, in this case each frame is subject to a procedure to validate the entrance.

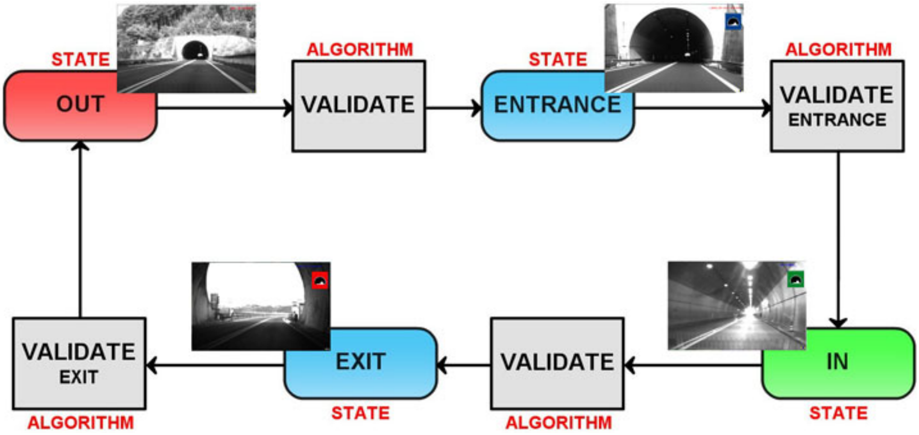


Fig. 3. State flow diagram of the system

If the system finds a positive entrance match, a new state is triggered and therefore the FSM enters the *ENTRANCE* state; the next state to be validated is the *IN* one, and so on.

This section describes algotithms and procedures used for the classification of each frame as shown in figure 3; the validate blocks are made up of a number of steps.

The validation procedures that are used for updating the FSM can be divided in two main classes:

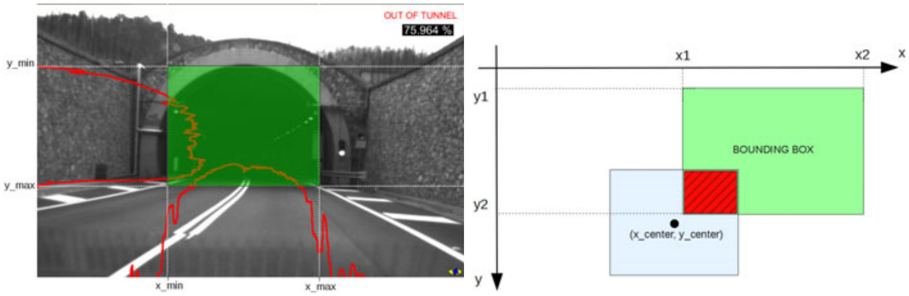
- *IN* and *OUT* validation: as explained in section 2.2, initially the system computes the horizontal and vertical histograms of dark pixels for the *OUT* state, or bright pixels for the *IN* state. Then the system validates the bounding boxes obtained by a histogram properties analysis.
- *ENTRANCE* and *EXIT* validation: in both transitions, the variance of histograms of black or white pixels is computed in an area of interest obtained from the previous state. For *EXIT* validation only, additional stages that involve image cropping and borders analysis are used. These two procedures are detailed in section 2.3 and 2.4.

2.2 Validation Algorithm for OUT and IN States

This section explains the algorithm performed in the *IN* and *OUT* states related to the *VALIDATE ALGORITHM* blocks in figure 3.

This procedure can be divided in three main steps:

- histograms analysis;
- bounding box construction;
- bounding box classification.



(a) (b)

Fig. 4. Bounding box construction and analysis. (a) Image showing the bounding box obtained from the analysis of the histogram of dark pixels. (b) Relationship between bounding box and its centered model.

Histogram analysis. In this step the algorithm builds the horizontal and the vertical histogram of the dark pixels, for the OUT state, and of the bright pixels for the IN state. These histograms give information about the numbers of the dark pixels, considering OUT state as example, for both rows and columns of the image. A dynamic threshold is used to select pixel values to be used for the histogram computation. This dynamic threshold is computed as follows:

$$Th = \begin{cases} th_{max} & if\ th \geq th_{max} \\ th_{min} & if\ th \leq th_{min} \\ value_{min} + P \cdot pixelValueSum & otherwise \end{cases} \tag{1}$$

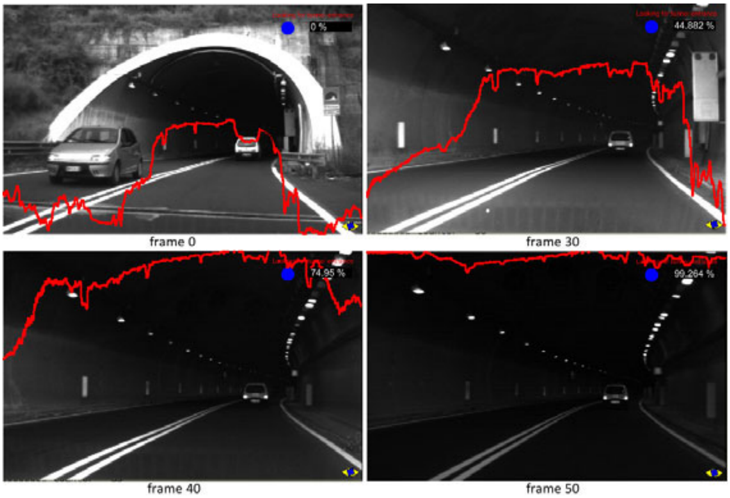
where th_{max} and th_{min} are saturation values, $value_{min}$ is the minimum value that can have a pixel and the $pixelValueSum$ is an experimental computed value that is weighted using a P weight computed as follows:

$$P = \frac{|brightness_{avg} - value_{min}|}{255} \qquad P \in [0, 1] \tag{2}$$

The P value is proportional to the difference between the darkest pixel ($value_{min}$) and the brightness average ($brightness_{avg}$). The value of th_{max}, th_{min}, $value_{min}$, $pixelValueSum$ are provided as input.

Bounding box construction. The interesting portion of both horizontal and vertical histograms are calculated from the maximum peak with a sliding window according to a predetermined threshold. This analysis returns a bounding box that identifies the darkest area of the image. Figure 4.a shows a bounding box built using the dark pixels histogram values analysis.

Bounding box classification. In this step information obtained from the previous step are analyzed in order to validate the bounding box as a tunnel entrance or exit.



frame 0

frame 30

frame 40

frame 50

Fig. 5. Histogram of the dark pixels values in *ENTRANCE* most significant frames

Classification is made in according to the following formula, that returns a validation percentage $P_{validate}$:

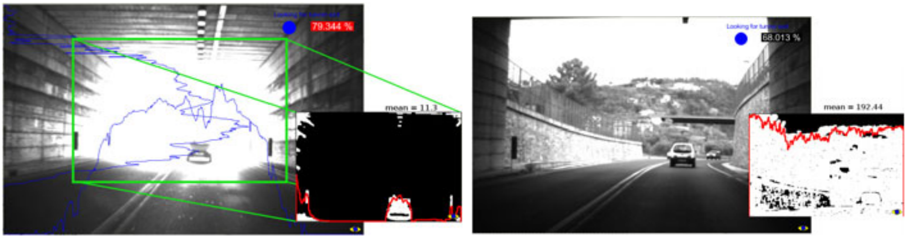
$$P_{Validate} = \frac{W_1 \cdot P_{dimension} + W_2 \cdot P_{center} + W_3 \cdot P_{filter}}{W_1 + W_2 + W_3} \tag{3}$$

Where $P_{dimension}$ is the ratio percentage between box and current image, P_{center} is the percentage of overlapping with a centered bounding box model as shown in figure 4.b, that provides a measure of how the bounding box in the current frame is centered. This value should increase as vehicle approaches the extremities of a tunnel. The weights W_x have been empirically computed. P_{filter} is related to box filtering functions that concern the aspect ratio of the tunnel extremities, the size of the bounding box, and a fixed percentage added to the box classification if the current frame has a size greater than a given percentage of the size of the box found in the previous frame.

As an example, the $P_{validate}$ computed on the image in figure 4.a is shown in the upper right corner. To fire the state transition from *OUT* or *IN* states to the *ENTRANCE* or *EXIT* states, the average value of the $P_{validate}$ values computed for the last n frames is used.

2.3 Validation Algorithm for ENTRANCE State

This section explains the *VALIDATE ENTRANCE ALGORITHM* block, that is used to fire the transition from *ENTRANCE* to *IN* states. As shown in figure 5, the dark pixels histogram saturates when the vehicle enters a



(a) (b)

Fig. 6. Exit from a tunnel. (a) Filtered and binarized bounding box before the tunnel exit, (b) filtered and binarized bounding box after the tunnel exit.

tunnel, this effect can be exploited to identify the transition. The variance increases when the vehicle is approaching a tunnel entrance and then decreases until it reaches a local minimum when the vehicle has just entered the tunnel. A parameter P that encodes the variance decrement is used to fire the transition. More precisely, P is computed as follows:

$$P = 1.0 - \frac{var}{varMax} \tag{4}$$

The value of P proportionally increases with respect to the decrement of the variance. P is compared against a threshold to fire the transition. Since the *ENTRANCE* has been activated, the variance value is computed, when it reaches the maximum, P start to increase.

Figure 7 shows the evolution of the variance in the entrance of a tunnel, the frame 0 indicates the activation of the *ENTRANCE* state; while the frame 40 represents the instant of the change of the state in which the vehicle is completely into the tunnel.

2.4 Validation Algorithm for EXIT State

In this step the procedure used to validate the exit from a tunnel is explained. This procedure is formed by two key point, an edge filtering procedure and the analysis of a histogram of the edges.

The bounding box obtained by the *IN* state validation block is filtered using a Sobel filter in order to extract the edges. In the transition from inside to outside the tunnel the presence of edges increases due to the change of camera controller parameters as shown in figure 6. The filtered box is binarized and a histogram of the edge pixels is computed for each box column. This histogram is averaged on the last n frames computing therefore the average of the number of edge pixels. The average is compared with a threshold to determine the state transition. Figure 8 shows the temporal histogram average in the *EXIT* state.

3 Results

The system was tested on a computer with the following features:

- Cpu: Intel(R) Core(TM)2 Quad CPU Q9550 @ 2.83GHz;
- Ram memory: 8 GiB DIMM DDR2 Synchronous 800 MHz (1.2 ns);
- Video Card: GeForce 9800 GTX/9800 GTX+ 512 MB memory.

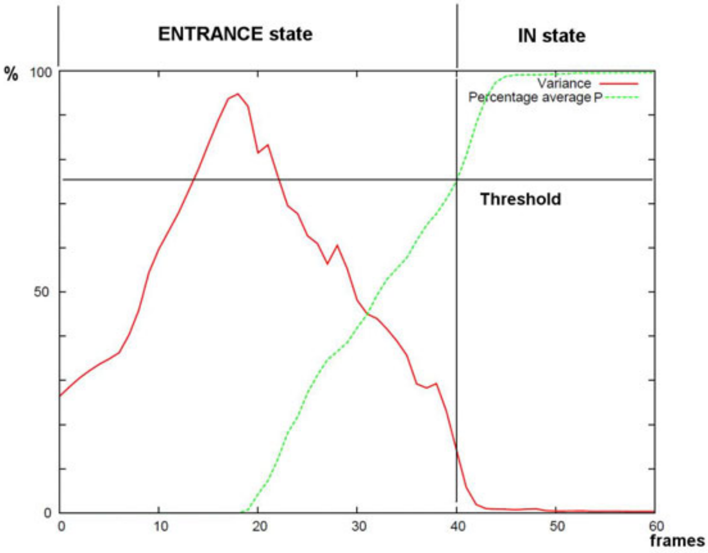


Fig. 7. Black pixels histogram variance and percentage average on frames in the *ENTRANCE* state

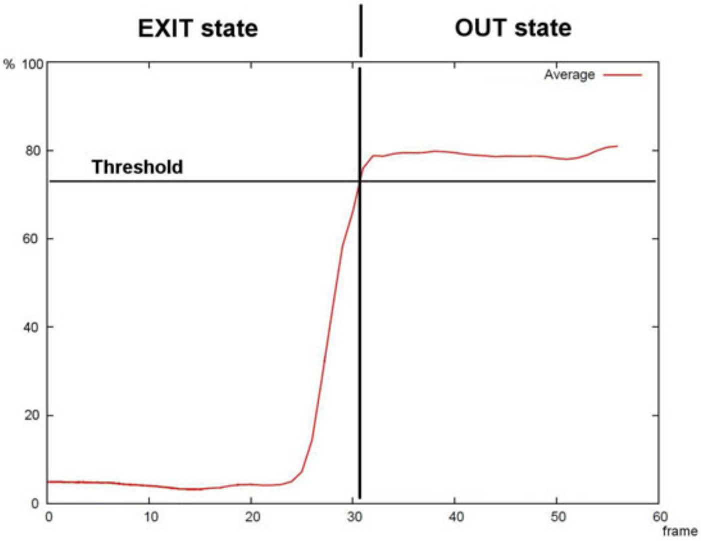


Fig. 8. Edge count average evolution from *EXIT* to *OUT* states

Table 1. Results of the algorithm

Sequence	% Correct Detection	Transitions	False Positives [frames]
first	100.0	80	13
second	83.3	48	0
third	100.0	32	350
total	94.4	160	363

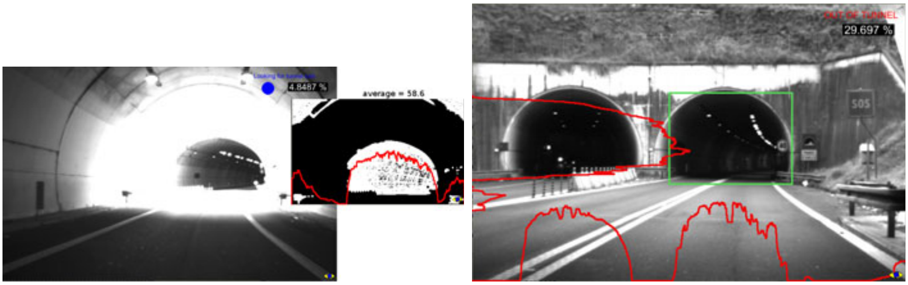
Three different sequences were used in the test, with 54686 frames in total at a frequency of 21 FPS. The image sequences were acquired during the day with frequent tunnels of varying length, with several entrances and exits, and normal traffic conditions. Tunnels have different shapes, with rectangular or arched entry and they are in urban, rural streets, and in highways. The images were acquired by a non calibrated camera with variable gain with a sensor in the visible spectrum with a NIR tail.

The system was proven to be able to detect the correct state with a total percentage of correct detection of 94.4% (see table 1). Missdetections appear in the case of urban settings. In this situation, the tunnel is just after a curve between the buildings of the city center, so the developed algorithm can not locate the entrance due to abrupt scenario variations.

In image sequences with the presence of trees along the road inducing sudden brightness changes, the system presented a limited number of false positives.

While it is actually not a missdetection, in 16.2% of the cases the tunnel exits are triggered in advance with respect to the ground truth due to strong light reflections produced by tunnel walls.

A further key parameter to judge the system performance is execution time. In fact, the system is meant to be run as a background process in conjunction with other ADAS functions. The application is able to process one frame at an average execution time of 750 μseconds, namely in less than 1 ms.



(a) (b)

Fig. 9. Problematic cases. (a) Exit from a tunnel: histogram mean are high because of the presence of another tunnel. (b) Twin galleries. The histogram have two relevant portions.

3.1 Problematic Cases

In this section two problematic cases are presented; this situation concerns consecutive road tunnels and twin frontal tunnels. In both situations system the behaves correctly and perform a correct tunnel detection.

Consecutive road tunnel. The presence of a tunnel entrance close after the exit of the current tunnel weakens the exit footprint in the image. This effect could lead to delayed detection in some cases. The above situation is shown in figure 9.a.

Twin tunnels. A frequent situation that can be especially encountered on motorways, is shown in figure 9.b. Two tunnels, one for each direction, create problems in identifying the histogram peak to detect the entrance bounding box. This problem has been solved giving priority to the central part of the histogram during the search for the largest portion, assuming that the tunnel of interest is in a central position and therefore in the central part of the image. The use of lane markings is also being considered.

4 Conclusions

The system for the recognition of road tunnels presented in this article has proven to work with a sufficiently high robustness.

The system was tested on different sequences including different tunnel situations: rural roads, highway tunnels, downtown with surrounding buildings, twin tunnels, consecutive tunnels, tunnels with several types of lighting or no lighting at all, tunnels with differently shaped entrances.

The application demonstrated to be able to recognize these tunnels and is not affected by traffic conditions at the tunnel entrance or exit.

The system can run in 750 μseconds and therefore is suitable to be used in conjunction with other ADAS systems.

References

1. Broggi, A., Mazzei, L., Porta, P.P.: Car-driver cooperation in future vehicles. In: Procs. Intl. Conf. on Models and Technologies for Intelligent Transportation Systems, Rome, Italy (June 2009)
2. Daytime Running Lights Deliverable 3: Final Report (October 2003)
3. D.M. 5 giugno,"Sicurezza nelle gallerie stradali". Pubblicato nella Gazzetta Ufficiale (217) (Settembre 18, 2001)
4. Ziemer, R.E., Peterson, R.W.: Introduction To Digital Communication, 2nd edn. Prentice-Hall, Englewood Cliffs (2000)
5. Cabani, I., Toulminet, G., Bensrhair, A.: Contrast-invariant Obstacle Detection System using Color Stereo Vision. In: 11th International IEEE Conference on Intelligent Transportation Systems, ITSC 2008, Beijing (October 2008)

A New Dissimilarity Measure for Clustering Seismic Signals

Francesco Benvegna[1], Antonino D'Alessando[3], Giosuè Lo Bosco[1],
Dario Luzio[2], Luca Pinello[1], and Domenico Tegolo[1]

[1] Dipartimento di Matematica e Informatica
via Archirafi 34, 90123 Palermo, Italy
[2] Dipartimento di Fisica e Chimica della terra
Via Archirafi 36, 90123 Palermo, Italy
[3] Istituto Nazionale di Geofisica e Vulcanologia
Centro Nazionale Terremoti, Italy
francesco.benvegna@unipa.it, antonino.dalessandro@ingv.it,
giosue.lobosco@unipa.it, dario.luzio@unipa.it, pinello@unipa.it,
domenico.tegolo@unipa.it

Abstract. Hypocenter and focal mechanism of an earthquake can be determined by the analysis of signals, named waveforms, related to the wave field produced and recorded by a seismic network. Assuming that waveform similarity implies the similarity of focal parameters, the analysis of those signals characterized by very similar shapes can be used to give important details about the physical phenomena which have generated an earthquake. Recent works have shown the effectiveness of cross-correlation and/or cross-spectral dissimilarities to identify clusters of seismic events. In this work we propose a new dissimilarity measure between seismic signals whose reliability has been tested on real seismic data by computing external and internal validation indices on the obtained clustering. Results show its superior quality in terms of cluster homogeneity and computational time with respect to the largely adopted cross correlation dissimilarity.

1 Introduction

In seismically active areas often occurred earthquakes that produce very similar waveforms (multiplets). A high level of similarity between the waveforms is a clear indication of events generated in a small seismogenetic volume, with similar source mechanisms. These events can be associated with both tectonic [1,2] and volcanic activity [4]. Based on the similarity between complete seismograms of microearthquakes occurred on the San Andreas Fault, Geller and Mueller deduced that their hypocenters can't be distant from each other by more than a quarter of the dominant wavelength [3].

The definition of the new dissimilarity was inspired by a simple observation: a seismic signal is characterized by the overlapping of several wave trains (seismic phases) which, because of their different travel path, arrive at the recording

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 434–443, 2011.

point at different times. They relate to both body waves and surface waves. The body waves concerning the irrotational component of the displacement field (P) propagate faster than those concerning the solenoidal one (S) and even more than the surface or guided waves.

The common hypocentral location methods, based on P and S phases arrival times inversion, are generally not accurate enough for a reliable relative location of very close hypocenters (hypocentral spacing much smaller than the typical distance between the stations of the seismic network) and modeling of the focal mechanisms distribution in the source region. To determine differential arrival times with high accuracy, techniques exploiting waveform similarities have been proposed [1,7].

The dissimilarity functions based on signal cross correlation have been used to measure the difference degree between seismic events [8,9] and to provide more precise estimates of the differences in arrival times of P and S phases of similar events [10,11]. A new challenge needs to identify, some groups containing similar signals with respect to a predetermined criterion, in a large set of three-component signals.

Clustering technique and the related algorithms can be adopted to face that challenge. In the general case, cluster analysis play a central role in the design of data analysis systems [12]. Moreover, clustering allows analysts to discover the nature of the data for further analysis. Dissimilarity (similarity) functions are a fundamental ingredient of clustering procedures, and their discrimination ability can be measured by means of clustering validation indices [13]. Clustering validation indices can be divided into internal and external ones: the former gives a reliable indication of how well a partitioning solution captures the inherent separation of the data into clusters, the latter measures how well a clustering solution agrees with the *gold solution* for a given data set [15]. A gold solution of a generic dataset, can be also inferred by analyzing the data, i.e., by the use of internal knowledge via data analysis tools such as clustering algorithms.

A basic consideration about cross-correlation and/or cross-spectral dissimilarities, is that they are effective in forming subsets of similar events just if earthquakes included in each set are very close in space, magnitude and focal parameters domains and the waveforms recorded have a good signal to noise ratio. Another of its drawback is the computation time, that will necessary affect the adopted clustering algorithm. This is a very important point since the development of dense seismic networks, with 3 components broadband sensors, permit to collect a lot of seismological data that should be processed by clustering techniques.

In this paper we propose a new dissimilarity measure able to catch difference in shapes between waveforms. It has been used in conjunction with a hierarchical clustering algorithm and applied to a dataset of earthquakes waveforms and to another dataset of signals generated by bursts, both recorded by an Ocean Bottom Seismometers with Hydrophone (OBS/H) deployed in the southern Tyrrhenian sea. We compared its discrimination ability with that of a cross correlation based dissimilarity. Results show the effectiveness of using the proposed dissimilarity, in terms of cluster homogeneity validation index and computational time.

2 Dissimilarities Definitions

In this section the two dissimilarity measures used in this work are described. The first one is the classical *cross correlation dissimilarity*, the other one is the new designed mesure called *cumulative shape dissimilarity.*

We recall that the *cross correlation* between two vectors x_1 and x_2, both of length n, is so defined

$$R_{x_1,x_2}(k) = \begin{cases} \sum_{i=0}^{n-k-1} (x_1(i+k) - \mu_{x_1}) \times (x_2(i) - \mu_{x_2}) & \text{if } k \geq 0 \\ R_{x_2,x_1}(-k) & \text{otherwise} \end{cases}$$

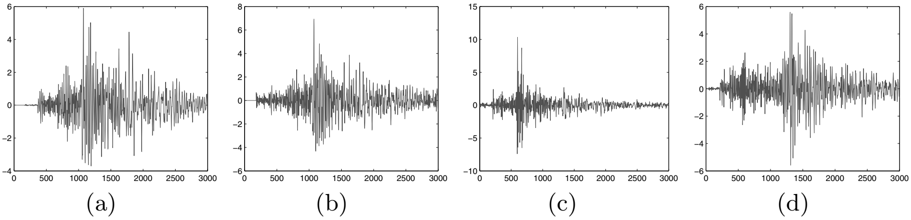
for $k = 1-n, .., n-1$, and where μ_{x_1} and μ_{x_2} indicate the means of x_1 and x_2 respectively. Consequently, the cross correlation dissimilarity between x_1 and x_2 is

$$\delta_R(x_1, x_2) = 1 - \frac{1}{\sigma_x \sigma_y} \underset{k=1,..,2n-1}{max} R_{x_1,x_2}(k-n). \tag{1}$$

Where σ_x and σ_y are the standard deviations of x_1 and x_2 respectively. Such dissimilarity is largely used to catch difference in shape between seismic signals, but in this context it has also shown some drawbacks. In fact, it is ineffective in forming subsets of similar events if earthquakes included in each set are not very close in space, magnitude and focal parameters domain, and noise is present in the recorded signal. Moreover, for a signal of length n its computational time is $O(n^2)$. The definition of the new dissimilarity was inspired by a simple observation: a seismic signal is characterized by two types of waves: body waves and surface waves. The body wave, especially the first P and S arrival times, are less sensitive to the travel path and clearly have no phase overlapping. Moreover, these seismic phases have often the better signal to noise ratio, so we can use them to discriminate one wave from the others. A seismic dataset is often a set of aligned (or not[1]) signals which contain the two types of body waves: P wave and S wave. Both waves have a magnitude peak with high energy. Consideration about the nature of the data, leads to state the main properties of a good dissimilarity measure for seismic signals :

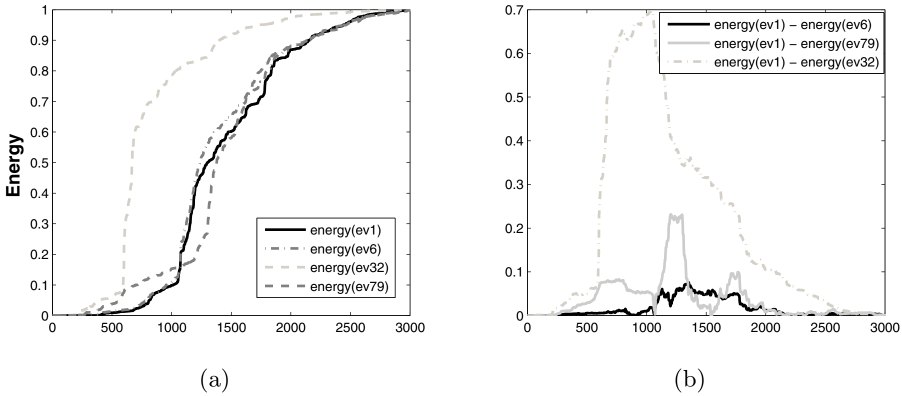
- it should give high weight to the difference among the initial part of the signals;
- it should be low sensitive to background and impulsive noise;
- it should be capable of detecting where two wave shapes are similar regardless of magnitude.

[1] Many technics are used to cut and to align the signals: a common phase is the pre-processing of the signal with denoising, P phase identification and cut.



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

Fig. 1. (a) event 1 (b) event 6 (c) event 32 (d) event 79



(a) (b)

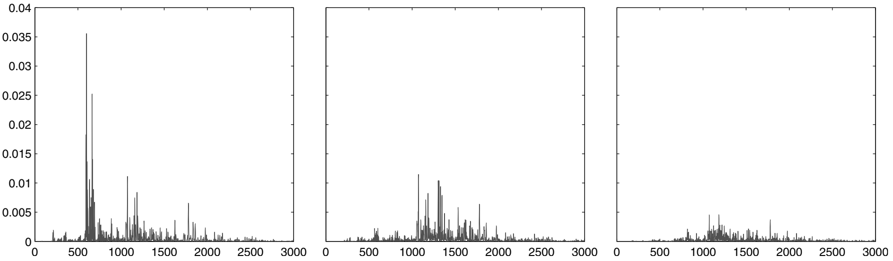
Fig. 2. (a) cumulative energy of the events; (b) difference between cumulative energies

The first two properties, can be satisfied by a dissimilarity acting on the cumulative energy of the signals rather than on their original waveforms. Of course, the peaks of the P wave and S wave are well visible on cumulative energy plot whereas the tail of the signal has a tiny impact. All the properties are finally satisfied by a dissimilarity that take into account the evaluation of the difference between cumulative energies.

Given two vectors x_1 and x_2 both of the same length n, and let s_1 and s_2 be their cumulative sums $s_i(k) = \frac{\sum_{r=1}^{k} x_i^2(r)}{\sum_{r=1}^{n} x_i^2(r)}$ $(i = 1, 2)$, we can calculate their absolute difference $sd(k) = |s_1(k) - s_2(k)|$. Finally, the new proposed dissimilarity, called *cumulative shape dissimilarity* δ_s is defined as:

$$\delta_s(x_1, x_2) = \sum_k \frac{|sd(k+1) - sd(k)|}{max_j\, |sd(j+1) - sd(j)|}. \tag{2}$$

Note that δ_s represents the sum of the derivative of the difference between the cumulative sums of x_1 and x_2. In figure 1 we report 4 examples of signal, in figure 2 their cumulative sums and the pairwise dissimilarities. Finally, in figure 3 we show the value of $|sd(i+1) - sd(i)|$ used to compute $\delta_s(x_1, x_2)$. Such example shows how similar shapes have lower dissimilarity values. It is important to note that the new measure δ_s have a remarkable computational time of $O(n)$.



(a) (b) (c)

Fig. 3. Derivative at sample point i of the difference between cumulative energies ($|sd(i+1) - sd(i)|$) (a) event 1 - event 32 (b) event 1 - event 79 (c) event 1 - event 6

3 Evaluation of a Dissimilarity Measure

In order to evaluate the performance of a dissimilarity, we have adopted three different indices. Two of them are related to the partitioning inducted by a clustering algorithm which make use of the dissimilarity, while the other one does not consider any partitioning information.

When using a dissimilarity measure in conjunction with a clustering algorithm, it is possible to evaluate its performance by means of *clustering internal and external indices*: the former gives a reliable indication of how well a partitioning solution captures the inherent separation of the data into clusters [15], the latter measures how well a clustering solution agrees with the *gold solution* for a given data set. A gold solutions for a dataset is a partition based on external knowledge of the data in classes, that can be also inferred by the use of internal knowledge via data analysis tools such as clustering algorithms. When the gold solution is not known, the internal criteria must give a reliable indication of how well a partitioning solution, and indirectly the used dissimilarity, captures the inherent separation of the data into clusters.

Let X a set of generic items $X = \{x_1, \ldots, x_N\}$, and $\mathcal{P} = \{p_1, \cdots, p_t\}$ a partitioning of X.

In our experiment we have adopted the `Homogeneity (H)` and `Separation (S)` internal indices [15] of a partitioning $\mathcal{P}$ produced by a clustering algorithm by using the dissimilarity δ, whose formulas are here reported:

$$H = \frac{1}{|X|} \sum_{i=1}^{t} \sum_{x \in p_i} 1 - \delta(x, \mu_i) \tag{3}$$

$$S = \frac{1}{\sum_{i \neq j} |p_i||p_j|} \sum_{i \neq j} |p_i||p_j| \delta(\mu_i, \mu_j) \tag{4}$$

where μ_i represent the centroid of a cluster p_i.

Note that both of the indices have to be considered: if $\forall x, y \; 0 \leq \delta(x, y) \leq 1$, they assume value in $[0, 1]$ and, the closer H and S are to 1, the better the partitioning of the data, and consequently the used dissimilarity.

When the gold solution is known, the so called external indices can be computed. Giving the partitioning $\mathcal{C} = \{c_1, \cdots, c_r\}$ corresponding to the gold solution for the dataset, an external index measures the level of agreement between $\mathcal{C}$ and $\mathcal{P}$. External indices are usually defined via a $r \times t$ contingency table T, where T_{ij} represents the number of items in both c_i and p_j, $1 \leq i \leq r$ and $1 \leq j \leq t$. For our experiment we have used the `Adjusted Rand index`[14].

$$R_A = \frac{\sum_{i,j} \binom{T_{ij}}{2} - \frac{[\sum_i \binom{T_{i.}}{2} \sum_j \binom{T_{.j}}{2}]}{\binom{N}{2}}}{\frac{1}{2}[\sum_i \binom{T_{i.}}{2} + \sum_j \binom{T_{.j}}{2}] - \frac{[\sum_i \binom{T_{i.}}{2} \sum_j \binom{T_{.j}}{2}]}{\binom{N}{2}}} \tag{5}$$

where $T_{i.} = |c_i|$ and $T_{.j} = |p_j|$. Also in this case, the closer R_A is to 1, the better the partitioning of the data, and consequently the used dissimilarity.

Besides the assessment of a dissimilarity function by making use of clustering validation indices, it is also possible to use an a priori information different form the gold solution. In the following, we will define a new index, called `Dissimilarity Optimality index` which make use of the sort of data items.

Let us assume now that X is a partially ordered set of generic items, whose sorting permutation $P = (i_1, i_2, \ldots, i_N)$ is known. In this case, the goodness of a generic dissimilarity δ on X can be established by comparing the sorting it induces on X with the sorting permutation P. In particular, what we expect from a good dissimilarity δ is that for each item x_i, its closest item with respect to δ is x_{i+k} with a small $|k| \geq 1$. The Dissimilarity Optimality index is so defined:

$$do = \sum_{i=1}^{n} \frac{|i - j - 1|}{N - 2} \text{ with } j = \underset{1 \leq k \leq N, k \neq i}{\operatorname{argmin}} \delta(x_i, x_k) \tag{6}$$

$do \approx 0$ is what we expect in case of good dissimilarity measure.

4 Experimental Results

On the 6th September 2002, at 01:21 UTC, a strong earthquake (M_W 5.9) occurred in the northern Sicilian offshore. The seismic event was recorded by the Istituto Nazionale di Geofisica e Vulcanologia (*INGV*) network and located at about 50 km in NNE direction, from the Palermo city. In the following months, more than a thousand of aftershocks were located in the same epicentral area [17]. In December 2009, to better monitoring the seismicity of the Palermo 2002 epicentral area, the Gibilmanna OBSLab of INGV installed an Ocean Bottom Seismometers with Hydrophone (*OBS/H*) near the epicentral area of the mainshock, at a depth of about 1500 m. The 3 Component velocity signals (Up-Down, Nord-Sud, East-West) was digitized with a 21 bit datalogger with a sampling frequency of 200 Hz. The OBS/H recorded several teleseismic and regional earthquakes and about 250 local micro-events not located by the on land network. The magnitude of the local events ranges between -0.5 and 2.5 M_L, and the delay between the S wave and P wave arrival times ($T_S - T_P$) ranges between

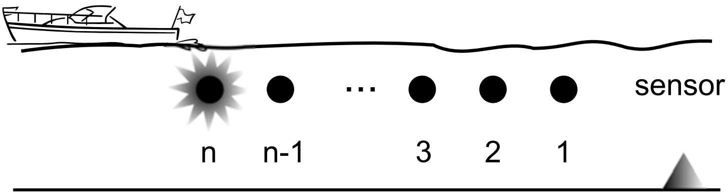


Fig. 4. Plan of the bursts experiment

0.2 s and 5 s. A visual analysis of the seismograms revealed some similarity. To better characterize the recorded micro-seismicity we located 159 micro-events, with Signal to Noise ratio greater than a selected threshold, with a 3C single station location technique based on the polarization analysis of the signals [16]. Among this microevents, 95 of them have been selected for our study. The resulting dataset, is denoted as *Palermo earthquake dataset*, and is finally composed by only the Up-Down component of 95 signals of length 3000 sample points.

Between April 7 and May 8 2010, was carried out a multidisciplinary geophysical investigation in the framework of the MEDOC project. In the first part of the experiment 4 wide angle seismic profiles, crossing the entire Tyrrhenian basin in East-West direction were acquired together with a fifth profile between southern Sardinia and Sicily. The seismic energy was produced by airgun bursts operating on the Sarmiento de Gamboa vessel, located at constant distance between them, placed at different distances from the OBS/H, and recorded with high signal to noise ratio. In particular, the airgun bursts occurs at regular interval times of 45s and the seismic sensor of the OBS/H records for each burst a signal s_i at time t_i that express the variation of the pressure level over time. Figure 4 shows the arrangement of the experiment. The acquired data define what is here named as *bursts dataset*, that can be considered a controlled dataset builded in order to have a well characterized set of signals to be used as a benchmark for problems involving seismic signals. The main assumption, is that close temporal explosions occurs at similar distances from the OBS/H. It is finally composed by the Up-Down component of 919 signals of maximum length 12000 sample points.

4.1 Results on Bursts Dataset

In order to test the relative merit of each distance over the bursts dataset we cutted the signals to a size useful to catch the meaningful part of the simulated burst. In particular we considered the first 1000 points of each signal because this part has an higher signal to noise ratio as explained in section 2. The performance of dissimilarities on this dataset has been measured by using the Dissimilarity Optimality index. This is due to the fact that the conducted experiment involves that signals recorded at closer instant times, should reveal similar shapes. The values of the distance optimality index for the cross correlation dissimilarity and the cumulative shape dissimilarity are 0.0033 and 0.0071 respectively.

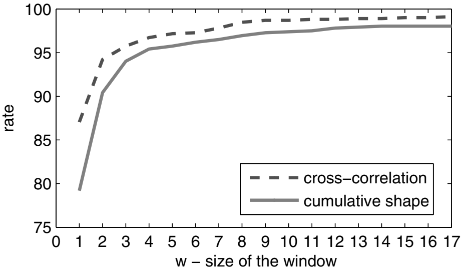


Fig. 5. Diagram of coverage proximity for w between 1 and 17

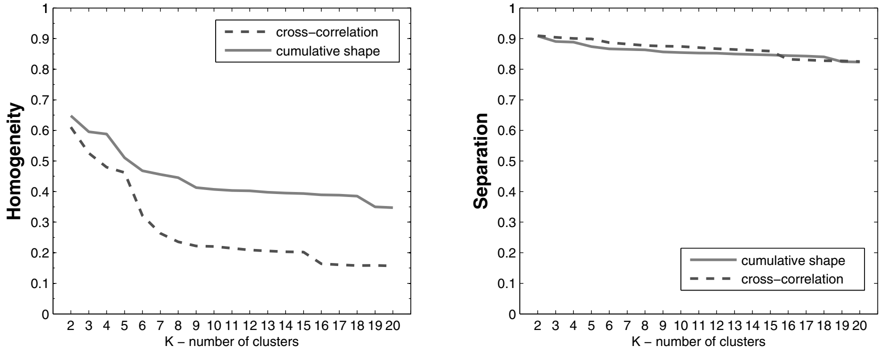
Both values are very close to 0 and their difference is very small.

We have also studied how the distance optimality index changes in terms of a temporal window w. In particular, for each signals x_i recorded at instant time t_i, we have computed the rate of how many times its closest signal x_j with $j = \underset{1 \leq k \leq N, k \neq i}{\text{argmin}} \; \delta(x_i, x_k)$ falls into a temporal window w, i.e $|t_j - t_i| \leq w$. We indicate this rate as *coverage proximity*. Figure 5 shows its computation for w ranging from 1 until 17.

The results (see figure 5) show that cumulative shapes have a coverage proximity of 80% vs 88% of cross correlation (8% difference) for $w = 1$. Anyway, this difference decreases very fast to 1% for $w > 1$. We can conclude that the performances of the two measures over the bursts dataset are almost equal.

4.2 Results on Palermo Earthquake Dataset

This dataset is composed by 95 signals of length 3000 sample points. The performance dissimilarities on this dataset has been measured by using the Homogeneity, Separation and Adjusted Rand indices. This is due to the fact that the we dispose of a gold solution established by the expert taking into consideration both its knowledge about the phenomena and the result of a hierarchical clustering algorithm using cross correlation dissimilarity. In particular, the spatial distribution of the hypocenters of the acquired data, suggests at least four well separated hypocenters clouds, close to the Palermo 2002 cluster [17]. This 4 clusters, had finally been splitted into 9 clusters with a variable number of events, by using the average link clustering algorithm in conjunction with the cross-correlation dissimilarity. The same clustering algorithm has been used to compute all the indices since it has been adopted by the expert to establish the gold solution. The first result is that the partitioning computed by the average link clustering in conjunction with the cumulative shape dissimilarity is perfectly equal to the gold solution (adjusted rand index equal to 1). Moreover, in order to better characterize this partitioning, we have computed its homogeneity and separation.



(a)

(b)

Fig. 6. Internal indices for the considered dissimilarities: (a) Homogeneity; (b) Separation

We report in figure 6(a,b) the homogeneity and separation indices of the two dissimilarities for different partitionings of K clusters ranging between 2 and 20.

The results show that the cumulative shape outperforms the cross-correlation in term of homogeneity and performs almost equally on separation.

5 Conclusion and Future Work

In this paper, a new dissimilarity measure between seismic signals called cumulative shape dissimilarity has been proposed. A number of tests have been done on two different dataset of earthquake events. The former is characterized by synthetic signal without gold solution in spite of the latter that, due to its real nature, have a gold solution proposed by an expert providing 9 cluster with a variable number of elements. Such datasets have been used to compare the cumulative shape dissimilarity with the cross correlation dissimilarity, that is actually largely adopted to differentiate waveforms in the context of seismic signals. In order to evaluate the goodness of the proposed measure, due to the heterogeneity of the two dataset, several indices have been considered (Dissimilarity Optimality, Homogeneity, Separation and Adjusted Rand). The test returns that the proposed measure have Dissimilarity Optimality and a Separation indices almost equal to the cross correlation ones, and a superior Homogeneity for all clusters values ranging from 2 to 20 (in average 1%). Anyway, the relevant difference has to be noted on the computational time, in particular cumulative shape measure is faster than cross-correlation ($O(n)$ vs $O(n^2)$). Future developments will be devoted to an extension of the cumulative shape on all the three-component signals, a new version taking into account weights for the signal samples, and to the study of the better conjunction between the new proposed dissimilarity and several kind of clustering algorithms.

References

1. Scherbaum, F., Wendler, J.: Cross spectral analysis of Swabian Jura (SW Germany) threecomponent microearthquake recordings. J. Geophys. 60, 157–166 (1986)
2. Console, R., Di Giovambattista, R.: Local earthquake relative location by digital records. Phys. Earth Planet. Inter. 47, 43–49 (1987)
3. Geller, R.J., Mueller, C.S.: Four similar earthquakes in central California. Geophys. Res. Lett. 7, 821–824 (1980)
4. Got, J.L., Frechet, M., Klein, F.W.: Deep fault plane geometry inferred from multiplet relative relocation beneath the south flank of Kilauea. J. Geophys. Res. 99(15), 375–386 (1994)
5. Aster, R.C., Scott, J.: Comprehensive Characterization of Waveform Similarity in Microeartquake data sets. Bulletin of the Seismological Society of America 83(4), 1307–1314 (1993)
6. Maurer, H.R., Deichmann, N.: Microearthquake cluster detection based on waveform similarities with an application to the western Swiss Alps. Geoph. J. Int. 123, 588–600 (1995)
7. Deichmann, N., Garcia-Fernandez, M.: Rupture geometry from high-precision relative hypocenter locations of microearthquake clusters. Geophys. J. Int. 110, 501–517 (1992)
8. Mezcua, J., Rueda, J.: Earthquake relative location based on waveform similarity. Tectonophysics 233, 253–263 (1994)
9. Menke, W.: Using waveform similarity to constrain earthquake locations. Bull. Seismol. Soc. Am. 89, 1143–1146 (1999)
10. Gillard, D., Rubin, A.M., Okubom, P.: Highly concentrated seismicity caused by deformation of Kilauea's depp magma system. Nature 384, 343–346 (1996)
11. Phillips, W.S., House, L.S., Feheler, J.: Detailed joint structure in a geothermal reservoir from studies of induced microearthquake studies. Journal of Geophysical Research 102, 745–763 (1997)
12. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. 31(3), 264–323 (1999)
13. Giancarlo, R., Lo Bosco, G., Pinello, L.: Distance Functions, Clustering Algorithms and Microarray Data Analysis. In: Blum, C., Battiti, R. (eds.) LION 4. LNCS, vol. 6073, pp. 125–138. Springer, Heidelberg (2010)
14. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification 2, 193–218 (1985)
15. Shamir, R., Sharan, R.: Algorithmic approaches to clustering gene expression data. Current Topics in Computational Biology, 269–299 (2002)
16. D'Alessandro, A., Luzio, D., D'Anna, G., Mangano, G., Panepinto, S.: Single station location of small-magnitude seismic events recorded by OBS in the Ionian Sea. Geophysical Research Abstracts, EGU General Assembly, Vienna, Austria, 12, EGU2010-8840 (2010)
17. Giunta, G., Luzio, D., Tondi, E., De Luca, L., Giorgiani, A., D'Anna, G., Renda, P., Cello, G., Nigro, F., Vitale, M.: The Palermo (Siciliy) seismic cluster of Septermber 2002, in the seismotectonic framework of the Tyrrhenian Sea-Sicily border area. Ann. of Geoph. 47(6), 1755–1770 (2004)

Character Segmentation for License Plate Recognition by K-Means Algorithm

Lihong Zheng[1] and Xiangjian He[2]

[1] School of Computing and Maths, Charles Sturt University, Australia
lzheng@csu.edu.au
[2] Faculty of E&IT, University of Technology, Sydney, Australia
xiangjian.he@uts.edu.au

Abstract. In this paper an improved K-means algorithm is presented to cut character out of the license plate images. Although there are many existing commercial LPR systems, with poor illumination conditions and moving vehicle the accuracy impaired. After examination and comparison of different image segmentation approaches, the K-means algorithm based method gave better image segmentation results. The K-means algorithm was modified by introducing automatic cluster number determination by filtering SIFT key points. After modification it efficiently detects the local maxima that represent different clusters in the image. The process is successful by getting a clean license plate image. While testing by the OCR software, the experimental results show a high accuracy of image segmentation and significantly higher recognition rate. The recognition rate increased from about 86.6% before our proposed process to about 94.03% after all unwanted non-character areas are removed. Hence, the overall recognition accuracy of LPR was improved.

Keywords: image segmentation, LPR, K-means algorithm.

1 Introduction

There are many problems with the License plate recognition (LPR) systems. In this paper we present a system based on K-means algorithm which efficiently improves the accuracy of character segmentation.

LPR is a system to identify a vehicle by recognizing the captured license plate image. It has been applied in numerous applications such as automatically identifying vehicles in a car park, and detecting and verifying stolen vehicles. There are many commercial LPRs around the world. Among these existing systems, there are two types of LPR. One uses commercial Optical Character Recognition (OCR) software to recognize the characters. The second kind uses learning based method to identify the characters. Both of them claim higher accuracy (99%) under controlled conditions and the cameras are mounted in fixed locations without mobility. However, with poor illumination conditions and moving vehicle the accuracy impaired. When using an OCR for character recognition, it is crucial to correctly remove the license plate boundaries after the step of license plate detection. No matter which OCR is used, the

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 444–453, 2011.

recognition accuracy will be significantly reduced if the characters are not properly segmented. Therefore the characters in a license plate need to be highlighted and separated from the background area to improve the recognition accuracy.

Image segmentation is one of the key processes in LPR. It partitions the image into some constituted parts so that each constituted part contains one character and can be extracted for further processing. Image segmentation can be approached from two different perspectives: contour based or region based approaches [1-2]. Contour based method is a gradient-based segmentation method. It attempts to find the edges or boundaries directly from their high gradient magnitudes. Edge or boundary based approaches [3-4], and active contour based methods [5], are some good examples. On the other hand, in the region based methods, the segmentation is usually based on discontinuity and similarity of the gray level of the image or other features, such as color, texture, shape, etc. However, most of these methods are not working well in some cases where in the captured video data, the characters are connected each other or with the boundary. And moreover, boundaries are often shown in similar pattern to the characters and sometimes boundaries are broken. Therefore, the recognition rate is very low due to poor segmentation of characters.

In this paper, we first present an overview of the different approaches in image segmentation. We then proposed the modified K-means algorithm in Section 4.3. It is applied in a real time LPR system. It can identify the character areas in a car license plate image and remove non character areas of license plates correctly and efficiently. The novelty of the proposed method is that the parameters in this algorithm are determined automatically in Section 4.3.1 and the optimal character region is founded by a smart searching in Section 4.3.2. Later, the experimental results verify that the proposed method is successful in terms of improving the recognition accuracy.

2 Review on Image Segmentation

Among various types of methods on image segmentation, two different types of image segmentation methods are contour based methods, and region methods. In this section, an overview of these two major different approaches on image segmentation is presented. The advantages and disadvantages of each type are discussed.

Firstly, a contour based method attempts to identify edge pixels and then link them together to form the required boundaries. The edge or boundary approach [3] is a gradient-based segmentation method. It attempts to find the edges directly from their high gradient magnitudes. The edge approach is similar to the boundary approach, and also uses gradient information. It measures the rate of change in a function such as the image brightness function. However, edge/no edge decision is made locally and prematurely, therefore random edge segments are found everywhere in the image [3].

Snakes, an active contour model [5], start with some initial boundaries represented in the form of spline curves, and iteratively modify them by applying various shrinking or expansion operations. They are autonomous and self-adapting in their search for a minimal energy state. They are relatively insensitive to noise and other ambiguities in the images because the integral operator is an inherent noise filter. They can be used to track dynamic objects in temporal as well as the spatial dimensions. However, they can often get stuck in local minima states. Their accuracy is governed by the convergence criteria used in the energy minimization technique; higher accuracies require tighter convergence criteria and hence, longer computation times. Therefore, time consuming and information loss are their main pitfalls.

A region based technique divides an image into smaller parts and merges neighboring areas that have the same features. Adjacent areas are merged according to some criteria such as homogeneity or sharpness of region boundaries. In the region approach, each pixel is assigned to a particular object or region. A good example of region based segmentation is the graph-based method [7]. Graph-based methods model the image as a weighted, undirected graph. Usually a pixel or a group of pixels are associated with nodes and edge weights define the similarity between the neighborhood pixels. The graph is then separated according to a certain design criterion to model "good" clusters. The process of dividing a graph is a recursive bipartition until some termination criterion is met. Often, the termination criterion is based on the same cost function. Some popular algorithms of this category are normalized cuts [8], random walker [9], minimum cut [7], minimum spanning tree-based segmentation [10], and ratio cut [11]. These researchers are seeking different ways to generalize and perform efficient iterated region based segmentation.

K-means algorithm [12] is one kind of region based clustering methods. It aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. It is to find the centers of natural clusters in the data as well as in the iterative refinement approach. The details are discussed in Section 4.3.

In general, contour based technique relies on edge or boundaries detection. The need to connect together broken contour lines makes this technique prone to failure in the presence of blurring. Active contour based method is iterative searching and time consuming and information loss are its shortcoming. The region approach isolates objects resting on a contrasting background from the scene by using thresholds. Image was partitioned into smaller parts by merging the neighboring areas with the same features, such as color, text, homogeneity or sharpness. However, stringent restrictions on merging criteria create fragmentation; lenient ones overlook blurred boundaries and over merge [13]. After improvement K-means algorithm based method does efficiently segment the image into regions by knowing the optimal number of clusters.

3 License Plate Detection

As shown in [14], the basic idea of the detection algorithm was to use a variable scanning window moving around on the input vehicle image. The size of basic scanning window is set to be 48 × 16 and then scaled up to 300 × 100. At each position, the image area covered by the scanning window was classified using a pre-trained classifier as either a license-plate area (a positive decision) or a non-license-plate area (a negative decision). The classifier used in this algorithm was a significant extension of Viola and Jones' work [6] to license plate detection.

In our algorithm we construct a six-layer cascaded classifier to increase the detection speed, in which the first two layers are based on global features, Edge Density and Edge Density Variance defined in [14], and the last four layers are based on local Haar-like features. The classification process can be taken as a degenerate decision tree containing multi-layer classifiers as shown in Figure 1. A positive result from the first classifier triggers the evaluation of a second classifier. A positive result from the second classifier triggers a third classifier, and so on. A negative outcome at any layer leads to the immediate rejection of the image region (block). It is commonly seen that, for a given vehicle image, the majority of evaluated image regions are negative. Therefore, the cascaded classifier shown in Figure 1 attempts to reject as many negatives as possible at the earlier stages. As its consequence, this cascaded classifier leads to fast license plate detection.

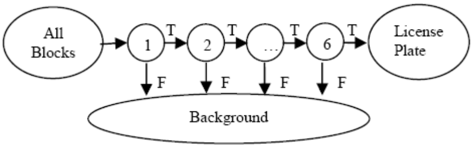
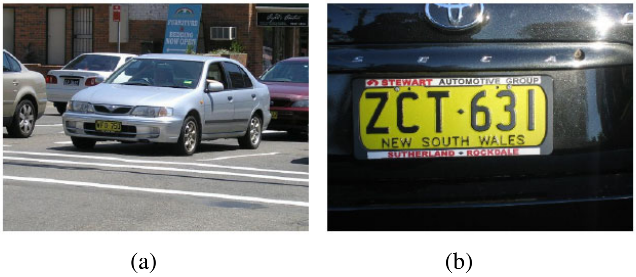


Fig. 1. Process of constructing a cascaded classifier [14]

To obtain the cascaded classifier which can make correct decisions, pre-classified positive samples (images containing license plates) and negative samples (images containing non-number-plates) are selected for training. In our experiments, all samples were obtained through manually labeling license plate areas in vehicle images captured under various conditions. The individual classifiers that together construct the cascaded classifier are trained independently.



(a) (b)

Fig. 2. Examples of small and large license plates

After the cascaded classifier has been trained and satisfies pre-defined classification accuracy, an input image is selected and a variable scanning window moves around the whole image space. To detect license plates of multiple sizes, the detection is carried out using multiple scales. Figure 2 gives an example that shows detectable license plates using our algorithm that have approximately the minimum size of 48 × 16 pixels (see Figure 2(a)) and maximum size of 300 × 100 pixels (see Figure 2(b)). In both cases, the vehicle images have 512 × 384 pixels.

4 Character Segmentation of License Plates

Once the license plate area is found, the following step is to find the characters contained in the license plate. The procedures of character segmentation include three different tasks. The first one is character height estimation. The upper and lower boundaries of character are located and used to obtain the character height. Then, we move to estimate character width. The final step is to cut character segments based on K-means algorithm.

4.1 Character Height Estimation of License Plates

This step contains three parts: colour reverse, vertical edge detection and horizontal projection histogram

4.1.1 Colour Reverse

The license plates in the New South Wales (NSW) State, Australia have many different formats, colours, and alignments. For instance, white in black, black in white, black in yellow, etc. are commonly used colour combinations. Colour reserve step is necessary before we assign right colour (i.e., black) to the characters of a license plate and hence obtain a correct binary image of the license plate. It makes the colour of the characters on a license plate be black. It is done based on a statistical analysis of edges. Given the located image, we pick *l* horizontal symmetrical lines on the license plate image. The colour index C_I is calculated as the average amount of the cross points (where pixel value changes between black and white) along each line in horizontal direction. Cross point number is increased by one if there is a foreground point.

$$C_I = \frac{1}{l}\sum_j \sum_i F(i,j) \tag{1}$$

where j is from 1 to *l* and *i* is from 1 to *N*, and *F* (*i, j*) is 1 if it is foreground point at location (*i, j*). If C_I value is over a statistical selected threshold, the candidate image is labeled as an image to be converted. Otherwise keep the original value for next step.

Therefore, all candidate license plates will look as black-in-white, and the consistency of license plate colour combination is made.

4.1.2 Vertical Edge Detection

Once a rectangular area of the license plate is detected and located, a simple area enlargement computation is applied so that the enlarged area will fully contain the license plate. Figure 3 shows some examples of final rectangular areas of the license plate images.

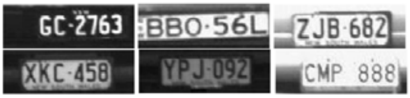
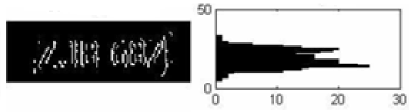


Fig. 3. Images Samples of Located License plates

Note that the located area shows stronger connectivity in vertical direction than horizontal direction, we perform vertical edge detection on the license plate images as shown in Figure 4 through the computation of horizontal gradient. At each pixel, use the Sobel mask of [-3 0 3; -10 0 10; -3 0 3] to compute the horizontal gradient value [15]. Then, use the Otsu [16] method for binarization to obtain the vertical maps. The edge pixels are represented using white pixels and other pixels using black pixels. Figure 4(a) shows the vertical edge pixels of the sample images.

4.1.3 Horizontal Projection Histogram

Although projection histogram is not a new concept, it is used here to find the upper and lower bounds of a license plate after the vertical edge are obtained. We perform a horizontal projection to find the top and bottom position of characters. When all values of histogram bins along all lines in the horizontal direction are computed, the horizontal projection histogram is obtained. The mean value of the histogram is then used as a threshold to figure out where the upper bound and lower bound are. The horizontal projection histograms of the three sample images are displayed in Figure 4(b). Finally, the distance between upper and lower boundaries is recorded as the height value of characters.



(a) (b)

Fig. 4. (a) Vertical edge maps of images in Figure 3, (b) Horizontal projection histogram

4.2 Character Width Estimation of License Plates

The middle area between the upper and lower bounds is recorded and considered for character segmentation on the license plate. Image binarization and vertical projection are two steps for segmentation here. Each segment here may contain one or two characters. Note that we use segmentation results to estimate the width of the characters on the license plate for the following processes. Each segment does not need to be accurate to contain exactly one character.

Image binarization highlights the pixels of interest and suppresses the background pixels. Otsu's method [16] is used to optimally classify all pixels with values above this threshold as white (255 grey value), and all other pixels as black (0 grey value). Figure 5 shows the results of binarization on the images in Figure 3 after cutting the upper and lower boundaries of the license plates.

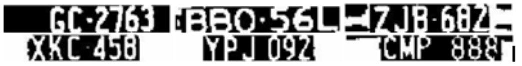


Fig. 5. Binarized images of the images in Figure 3 with upper and lower bounds removed

Similarly, we perform a vertical histogram projection to find the gaps between characters on a license plate. Each license plate is separated into blocks horizontally by the zero points in the projection histogram. Figure 6 shows the vertical projection of three images in Figure 5 and the segments (or blocks) of the images.

To estimate the width of a character after the segmentation process above, we take into account only widths of all blocks except the two smallest blocks and the two biggest ones. The averaged widths of these blocks are used as the estimated width of characters. The estimated height and width of characters will be used in the following step for character extraction.

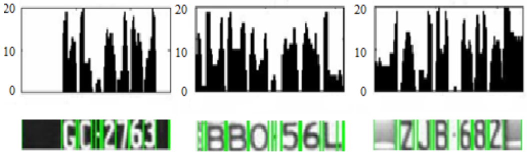


Fig. 6. Vertical projection of the images in Figure 5 and character segments

4.3 K-Means Algorithm Based Accurate Character Segmentation

4.3.1 K-Means Algorithm

The K-means [12] algorithm was invented in 1956. It is an iterative technique that is used to partition an image into *K* clusters. The basic idea is as follows:

1) Pick *K* cluster centers randomly,
2) Assign each pixel in the image to the nearest cluster based upon the similarity parameter such as Euclidean distance of intensity,
3) Recalculate the cluster centers for the new clusters by averaging all of the pixels in the cluster,
4) Repeat steps 2 and 3 until convergence is reached (e.g. the pixels no longer switch clusters).

K-Means algorithm typically converges to a solution very quickly as opposed to other clustering algorithms. But it may not return the optimal solution. A drawback of this algorithm is that the quality of the solution depends largely on the initial set of clusters and the value of *K*. An inappropriate choice of *K* may yield poor results. The optimal *K* value is identified automatically by SIFT based method described in the following section. The experimental results are shown in Figure 7 below.

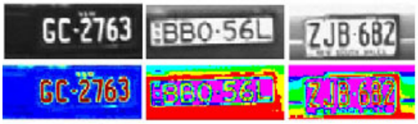


Fig. 7. Segmented areas in different colors of license plate sample images

4.3.2 Selection of K by SIFT Algorithm

Scale Invariant Feature Transform (SIFT) algorithm developed by Lowe David [17], is currently one of the best ways to find the invariant features. The SIFT features are proven that to be invariant to image scale and rotation, and are shown to provide robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination. SIFT combines a scale invariant region detector using the Difference of Gaussian (DoG) and a descriptor based on gradient distribution in the detected regions. Firstly, local peaks (termed key-points) in a series of difference-of-Gaussian (DoG) images are found. Then, candidate key-points are located based on the measures of their stability. After filtering of SIFT key points make a good auto determination of cluster *K* value for K-means algorithm. Figure 8 below show the examples where SIFT key points are.

Once these extremely discriminative SIFT features are obtained, it is important to have a prior that obtained in previous steps. After filtering some key points leading to more clusters, the optimal cluster *K* value of the K-means algorithm turn out depend on the choice of prior knowledge such as character bounds.

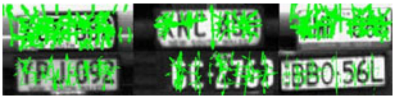


Fig. 8. Key points identified by SIFT algorithm

5 Experiential Results and Comparison

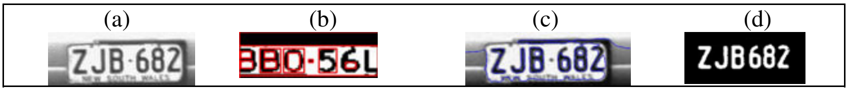
To test the recognition algorithm, we apply the proposed algorithm on 587 license plate images with 3502 characters located by the license plate detection step. All characters are correctly segmented. The false positives (i.e., non-character areas that are segmented) are mainly due to the false detection of license plates. Only 7 out of total 594 license plate detected regions are non-license plates. So our correct segmentation rate is still very high and is about 98.82% even taking into account the rate of wrongly detected license plate regions. The binary enlarged license plate images are obtained by our proposed method and are sent to the OCR software (Tesseract) for recognition. With slight modification of the OCR code, all segmented characters on the license plates are recognized correctly. A performance comparison with various methods is shown in Table I. Our license plate detection rate is 96.4%.

Table 1. Performance Comparison

References	Detection Accuracy	Segmentation/Recognition Accuracy
[18]	93.2%	95%
[19]	97.1%	96.4%
Proposed	96.4%	98.82%

Furthermore, an empirical evaluation of the proposed segmentation method(s) and other segmentation methods such as traditional CCA and active contour (AC) based method [5] is presented.

Firstly, traditional CCA uses the connectivity between a pixel and its neighbors to label the pixel and merge into same group. Given an image, CCA assigns labels to a pixel such that adjacent pixels of the same features are assigned the same label. The image is copied into a small or large array. It does multiple scans to label the pixels as belonging to one of many different groups. However, if characters are connected each other, it is hard to separate them individually. Secondly, active contour based approach starts with some initial boundaries and iteratively modify them by applying some shrinking or expansion operations. It gets stuck in local minima states due to overlook minute features in the process of minimizing the energy over the entire path of their contours. It is a time consuming method. It takes more than eleven times of the time that CCA takes and ten times of the time that our approach uses. As shown in Table II, our approach found the best interested area of clear characters finally.



Notes: (a) Original image (b) CCA method (c) Active contour based method (d) Our approach

Fig. 9. Comparison of original license plate with images developed using other approaches

Table 2. Comparison of experimental results with images developed using three alternative approaches

Accuracy rate of LPR	**Methods**		
	CCA	***AC***	***Ours***
Detection rate of license plates	96.4%	96.4%	96.4%
Accuracy of character segmentation	91.0%	76.1%	98.82%
Average time of segmentation (s)	0.154	2.21	0.204
Character recognition rate	98.7%	98.7%	98.7%
Overall recognition rate	86.6%	71.9%	94.03%

6 Conclusions

In this paper, we have constructed a cascaded classifier consisting of 6 layers for license plate detection using both global edge features and local Harr-like features. The classifiers on the first two layers exclude more than 80% non-plate regions from further training or testing and hence greatly increase the detection speed in the next four layers. The classifiers on the next four layers are based on local Haar-like features. With a small number of features, we can obtain very high detection rate with very low false positive rate even when the license plate detection algorithm is used under various complex environments. A real-time detection speed is achieved.

Moreover, we have proposed a method to segment the characters of car license plates. This is a crucial work after the detection of license plates and before the use of OCRs for character recognition. The process is successful through various techniques including image binarization, vertical edge detection, horizontal and vertical image projections, and modified K-means segmentation algorithm. Various well-known techniques are applied to come out with the innovative algorithm in this paper. The Tesseract OCR software was used to test our results. The experimental results show a significantly higher recognition rate 94.03% after character is segmented. Hence, the overall recognition accuracy has been improved.

References

1. Shapiro, L.G., Stockman, G.C.: Computer Vision. Prentice Hall, Englewood Cliffs (2002)
2. Achanta, R., Estrada, F., Wils, P., Süsstrunk, S.: Salient Region Detection and Segmentation. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) ICVS 2008. LNCS, vol. 5008, pp. 66–75. Springer, Heidelberg (2008)
3. Malik, J., Belongie, S., Shi, J., Leung, T.: Contour and Texture Analysis for Image Segmentation. International Journal of Computer Vision 43(1), 7–27 (2001)
4. Ko, B.C., Nam, J.Y.: Object-of-interest Image Segmentation based on Human Attention and Semantic Region Clustering. Journal of Optical Society of America A 23(10), 2462–2470 (2006)
5. Chan, T.F., Vese, L.A.: Active Contours without Edges. IEEE Transactions on Image Processing 10(2), 266–277 (2001)
6. Viola, P., Jones, M.J.: Robust Real-Time Face Detection. International Journal of Computer Vision 57(2), 137–154 (2004)
7. Shi, J., Malik, J.: Normalized Cuts and Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 888–905 (2000)
8. Jermyn, I.H., Ishikawa, H.: Globally Optimal Regions and Boundaries as Minimum Ratio Cycles. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(10), 1075–1088 (2001)
9. Sarkar, S., Soundararajan, P.: Supervised Learning of Large Perceptual Organization: Graph Spectral Partitioning and Learning Automata. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(5), 504–525 (2000)
10. Soundararajan, P., Sarkar, S.: Analysis of Mincut, Average cut, and Normalized Cut Measures. In: Third Workshop Perceptual Organization in Computer Vision (2001)
11. Wang, S., Siskind, J.M.: Image Segmentation with Ratio Cut. IEEE Transactions on Pattern Analysis and Machine Intelligence 25, 675–690 (2003)
12. Kanungo, T., et al.: An Efficient K-means Clustering Algorithm: Analysis and Implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 881–892 (2002)
13. Chazelle, B.: Application challenges to computational geometry: CG Impact Task Force Report. Technical Report TR-521-96, Princeton University (1996)
14. Zhang, H., Jia, W., He, X., Wu, Q.: Learning-based License Plate Detection in Vehicle Image Database. International Journal of Intelligent Information and Database Systems (IJIIDS), Inderscience 1(2), 228–243 (2007); ISSN: 1751-5858
15. `http://www.pages.drexel.edu/~weg22/edge.html`
16. Otsu, N.: A Threshold Selection Method from Gray-level Histograms. IEEE Transactions on Systems, Man, and Cybernetics 9, 62–66 (1979)
17. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
18. Wu, C., et al.: A Macao License Plate Recognition System. In: IEEE Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, vol. 7, pp. 4506–4510 (2005)
19. Guo, J.M., et al.: License Plate Localization and Character Segmentation with Feedback Self-learning and Hybrid-binarization Techniques. In: IEEE Region 10 Conference, pp. 1–4 (2007)

A Video Grammar-Based Approach for TV News Localization and Intra-structure Identification in TV Streams

Tarek Zlitni, Walid Mahdi, and Hanène Ben-Abdallah

MIRACL, Multimedia, Information Systems and Advanced Computing Laboratory
University of Sfax, Tunisia
{tarek.zlitni,walid.mahdi}@isimsf.rnu.tn,
hanene.benabdallah@fsegs.rnu.tn

Abstract. The growing number of TV channels led to an expansion of the mass of video documents produced and broadcast on TV channels according to precise rules (e.g. consideration of the graphic charter, recurring of studios…). Thus, the use of a priori knowledge deduced from these rules contributes to the amelioration of the quality of segmentation and indexing of video documents. However, the effectiveness of automatic video segmentation works depends on video type. So, for a better quality of the segmentation, it is necessary to consider a priori knowledge concerning video types. In this context, this paper suggests an approach based on video grammar to identify programs in TV streams and deduce their internal structure. This approach attempts to automatically extract a priori knowledge to conceive the grammar descriptors. The study case of TV news programs is selected to validate the adopted approach since it is one of the most important types of multimedia content.

Keywords: TV programs localization, TV news structuring, video indexing, video grammar, a priori Knowledge.

1 Introduction

Given the increasing number of TV channels, a smart access to their broadcast contents represents a real challenge. The diffusion generates opaque streams whose duration exceeds several hours. An efficient multimedia content structuring approach should first of all proceed by the identification of a program in a large stream, and then detect all various units making up this program. According to production rules, the location indices or separations between the internal entities are recurring (jingles, studio decors). This work aims to use the recurrence of these indices in a process of inter-segmentation and internal structuring of the audiovisual contents.

A priori knowledge is extracted and modeled as descriptors and stored for future uses. These descriptors are used to generate visual grammars which store this knowledge in a structured and relevant way. As a result, each TV channel has its own grammar based on the recurring and discriminating descriptors for the automatic content structuring broadcast. To highlight the concept on which the grammar generation is based, TV news programs are selected as a representative study case for this work.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 454–463, 2011.

The remainder of this paper is structured as follows. The second section shows the motivation of this work and explains its general concept. In the third section, the steps of extraction of a priori knowledge and the coding of the relevant descriptors are presented. The fourth section deals with descriptors extraction, and the fifth section explains the structuring and storing of the obtained descriptors. Finally, conclusion and some future research directions are given in section six.

2 Motivation and Grammar Concept

TV channels broadcast several hours of various programs in a continuous way. Although resulting streams are long and heterogeneous, the points of location between different programs and the units within each one are recurring for all program instances. For example, the start jingles are quasi-invariant for at least one year, and thus it would be unimportant to repeat the treatment of their detection to each program occurrence during this period. Based on these issues, this work intends to factorize the recurrence of the delimitation points and to index them in a structured way as a grammar. The extraction, modelling and storing of the descriptors of these points are used as a priori knowledge. They are subsequently exploited in the processes of segmentation as ways for detection and validation of the appearance of their descriptors.

This work provides a grammar which collects and structures a set of a priori knowledge useful for the identification and the structuring of TV programs (Fig. 1). The use of this knowledge has various advantages. On the one hand, it presents the

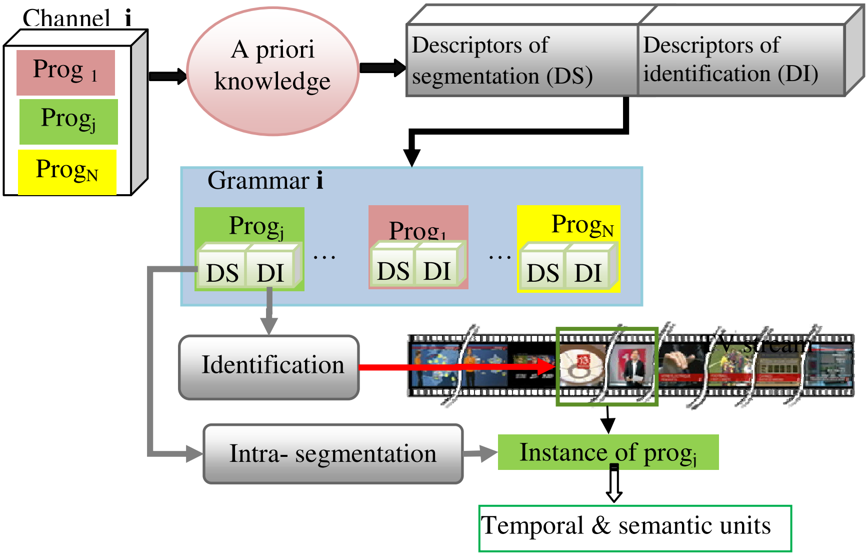


Fig. 1. Generation and exploitation of video grammar for the structuring of TV streams

principle of the descriptors factorization. In fact, video grammar is conceived mainly to model the recurring entities of TV channels. In other words, since the instance descriptors are always the same, they are extracted only once and reused with each process of segmentation (actually for a long duration). Such a step is undertaken in order to avoid tedious processes for the extraction of the primitives, thus the important profit in terms of execution (time calculation, memory capacity…). On the other hand, grammar also represents a complementary source to confirm the detected structural units. Consequently, a temporal or semantic unit is validated at the same time by the descriptors of signal level and the grammatical rules defined for the semantic concept of which this unit has been instanced. For example, a unit of the shot presenter type is detected by the appearance of a person shot and validated by the descriptors of this concept (presenter face and studio decor) defined by the grammar.

This way, each channel has its own grammar consisting of identification indicators of the programs in streams and the descriptors that play the role of indices for the deduction of the internal structure of the channel's programs.

3 A Priori Knowledge Extraction and Modeling

Although several previous works dealt with the segmentation of TV programs and TV news in particular [8], [10], few works start investigation from the inter-segmentation phase (identification) of a program in TV streams. This work highlights the identification phase and the location of the programs in streams since in the majority of the cases, the programs to be structured are incorporated in TV streams of long durations that can reach 24h and even more.

Once the programs are identified, the second part consists of enriching grammar by indices helping afterwards to identify the internal structure of these programs (1).

$$grammar(i) = \bigcup_{Prog_j \in Channel(i)} descriptors(Prog_j) \tag{1}$$

TV channel grammars consist of two types of descriptors (2) respectively for the two segmentation levels of a TV stream: inter-segmentation (identification) and intra-segmentation (internal segmentation) (Fig. 2) for each program $\boldsymbol{prog_j}$.

$$Descriptors(Prog_j) = DI(Prog_j) \cup DS(Prog_j) \tag{2}$$

- *DI* : identification descriptors.
- *DS:* internal segmentation descriptors.

The first type is made up of a priori knowledge used for the identification of the programs in video streams of a channel.

The second type consists of descriptors facilitating the internal structuring of the programs in a subsequent step. These descriptors are generally visual primitives which describe occurrences of the semantic points of anchoring (semantic entities) that delimit the internal units of a program.

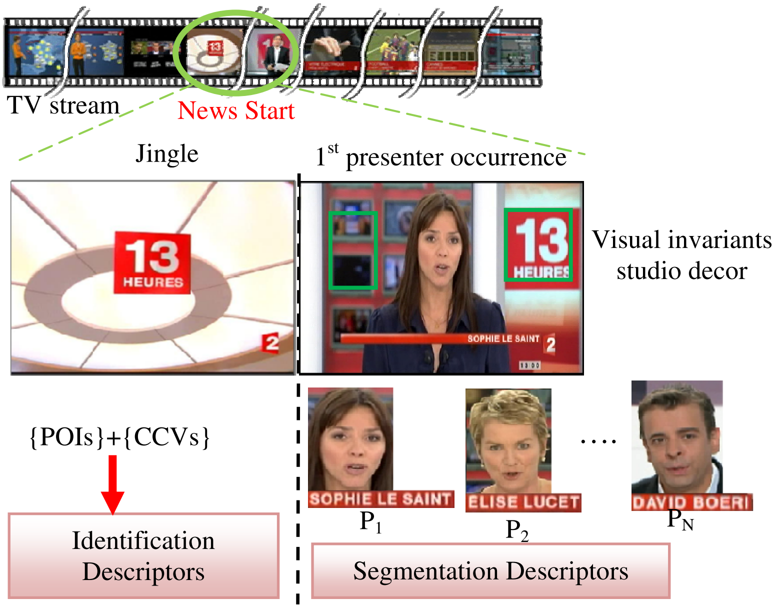


Fig. 2. Sample of the suitable descriptors for the identification and intra-segmentation of TV news

For TV news videos and according to literature [8], the apparition of the presenter(s) in the studio is considered as the anchoring point which delimits the temporal units of news programs, i.e. subjects which make up an instance of TV news program.

So, it would be essential to extract structure and store this knowledge in a relevant way in order to improve the semantic segmentation of the programs. Thus, for this programs type, in addition to the descriptors of trucking, grammar is supplied by visual studio decor descriptors and presenters' descriptors.

4 Descriptors Extraction

4.1 Identification Descriptors

The identification phase is based on Zlitni et al. [9] work's that dealt with the extraction of the adequate descriptors to distinguish the visual jingles from the various programs in a TV stream. Considering their specificity, two discriminative descriptors of video segments representing these jingles were chosen [4].

The first descriptor is a Point Of Interest descriptor (POI) chosen for the following reasons. POI is a point in an image which has particular properties; the peak is located where photometric information is most important within an image. These points are characterized by the robustness face to luminosity variations, blur effects and geometrical transformations.

The second descriptor is a colorimetric descriptor called Color Coherent Vector (*CCV*). Indeed, this descriptor represents each color by two values α and β which represent, respectively the number of coherent pixels and that of the incoherent ones. Moreover, the method used consists of classifying the pixels of the same color, in two categories: coherent and incoherent according to the size of the areas of the image to which they belong. Contrary to histograms, this descriptor presents the distribution of the colors considering their space dispersion.

4.2 Intra-segmentation Descriptors: Case of TV News

The intra-segmentation descriptors are indicators of unit change or the occurrence of an important event in a TV program (a goal in a sport program, new topic in TV news, etc). For the case of TV news, the presenter shot generally serves as an indicator of topic change. So, the descriptors of this entity are extracted and modeled (Fig. 3). Two characteristics specify this entity: presenter face and studio decor features.

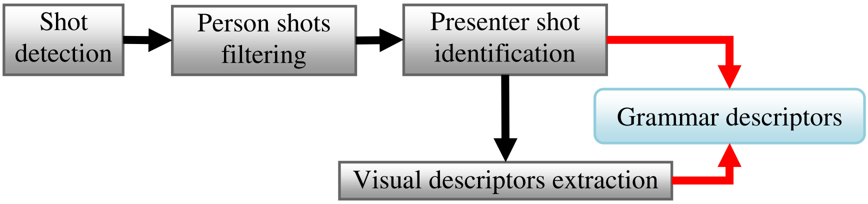


Fig. 3. Alimentation process of the grammar descriptors

4.2.1 Presenter Detection

Similarly to the graphic effects and studio decor of the programs which are invariant (at least for a long duration), the number of presenters per channel and type of program is quite limited. As for the news programs, the same persons are held in turn to present daily TV news programs. Based on this observation, it would be interesting to launch the process of detection and identification of some iteration and store their descriptors rather than call upon this process for hundreds of times during a few weeks.

Following the detection of start jingles the algorithm of identification of the presenter shot is launched. This algorithm is made up of two steps, (i) the detection and filtering of the person shots, (ii) the identification of the presenter's face.

a) Person shots detection and filtering

For the detection of the person shots, two phases are established: the shot detection [7] and the face detection.

To validate the person shots, the face detection technique presented by Viola and Jones [5] and developed by Lienhart and Maydt [6] is adopted, based on descriptors in cascades containing the wavelet of Haar. With this technique, the checking of the face presence in a zone of the image is based on the checking of the existence of a set of classifiers called characteristics of Haar. The application of these classifiers is achieved in cascade where the order of the classifier depends on its weight.

For the filtering of the presenter shots, a preliminary filtering of the person shots was initially carried out. These are shots containing at least a person. Since a person is identified by his face, the detection of this type of shots is based on the detection of faces. A person shot is then the shot where face(s) appear.

A shot is considered a presenter shot only if it satisfies two essential conditions inspired from the rules of production of TV news defined by Zlitni et al. [11].

- The existence of only one person in the shot (maximum two): there are one or two presenters per news program, generally localized in the shot center (Fig. 5).
- Front view persons: since the presenter addresses the viewers.

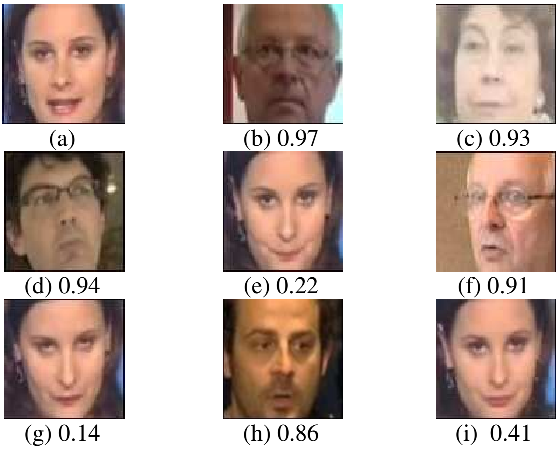
b) Presenter identification

In order to recognize presenter face among all the faces detected in TV news, the Eigenfaces approach was adopted: a face recognition approach[2].Eigenfaces consists in measuring the similarity of a requested image with the basic images. Each face image is regarded as a vector in a space having as many dimensions as pixels in the image. The characteristics of the image are extracted by a mathematical method of dimensionality reduction based on the principal components analysis (PCA). An adaptation of this approach consists in computing all the similarities of the faces (V) with the remainder (Fig. 4). It is an operation of clustering of the similar faces. Each face will have a set of similarities (3).

$$Sim(v) = Card(x|\, sim(v,x) \leq TH_{SIM}, \forall x \in V) \tag{3}$$

Since the presenter is the person who appears more in TV news, the face having the maximum of similarity will be considered the presenter face (4).

$$v(P_i) = \max\,(\{\mathrm{sim}(v)\}) \tag{4}$$



(a) (b) 0.97 (c) 0.93

(d) 0.94 (e) 0.22 (f) 0.91

(g) 0.14 (h) 0.86 (i) 0.41

Fig. 4. Distances similarity of face (a) with a set of faces appearing on TV news

An experimental study on different TV news programs was established to evaluate the method suggested for the presenter identification. Very interesting rates were raised (recall ≈ 95) (precision ≈ 92). With the detection phase, the rate of recall is taken into account. Indeed, even if there are false detections (not faces) they have similarity scores equal to zero and are isolated afterwards in the phase of filtering.

Each presenter occurrence is indicated in the grammar by two properties (5): descriptors of his face and his name.

$$P_k = P(f_k, n_k) \tag{5}$$

4.2.2 Decor Descriptors

The second descriptor to characterize the presenter shot is the decor consisting in the decoration of the plateau and the possible graphic effects encrusted in this shot.To represent the graphic recurrence of the decors, a set of visual invariants are selected to graphically symbolize the decor.

As already mentioned, the presenter appears in a studio with a invariant decor. Thus, to represent this invariance, specific zones are selected to extract the descriptors automatically. These zones are in the areas surrounding the presenter. Indeed, following the localization of the presenter's face, we deduce the invariants of the decor close to the face. To reduce description, we were limit to the two areas on the left and on the right of the face (Fig. 5).

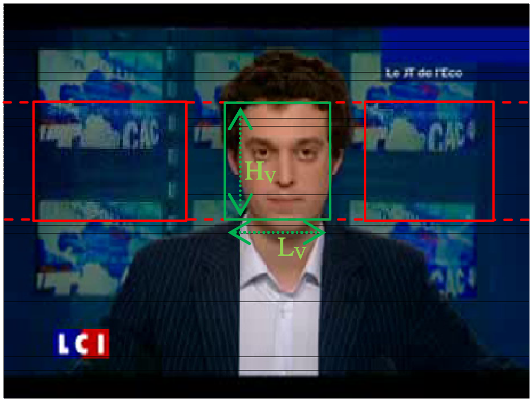


Fig. 5. Automatic localization and extraction of the invariant zones

The decor is usually characterized by particular textures visually easy to identify. The descriptors of texture have the property of identifying various areas in an image. They contain information about the space or statistical distribution of the colors. To compute these descriptors, the co-occurrence matrices are used [3]. These matrices measure the probability of appearance of the pixel pairs located at a distance δ in the image. They are based on the probability compute $M_{\delta,\theta}[i,j]$ which represents the number of times where a pixel of color level appears at a relative distance $\boldsymbol{\delta}$ of a pixel of level of color and according to a given orientation θ . For reasons of simplification of the descriptors complexity, the direction parameter was eliminated by fixing it at 0. Several second order metrics can be deduced from these matrices to characterize texture [1]. According to their discriminative effects for the various instances of the invariants, four metrics are retained: contrast (6), local homogeneity (7), homogeneity (8), and entropy (9).

$$C(k,n) = \sum_i \sum_j (i-j)^k \times M_\delta[i,j]^n \tag{6}$$

$$LHo = \sum_i \sum_j \frac{M_\delta[i,j]}{1+|i-j|} \tag{7}$$

$$Ho = \sum_i \sum_j M_\delta[i,j]^2 \tag{8}$$

$$E = \sum_i \sum_j M_\delta[i,j] \times ln(M_\delta[i,j]) \tag{9}$$

The similarity measured between areas which belong to the grammar and which are detected from the video is logically based on a function with a distance ***D*** calculated from a set of vectors f_i. Each vector includes the descriptors f_i corresponding to a specific area. Generally, this function is susceptible to noise and depends on the relevance of descriptors compared to a zone. To avoid this problem, the solution is to define a class for each zone covering a number of samples (i.e. several V_{fi} vectors) from the same spatial area but located on different images of the presenter shot as a first. Each class is considered as training data for a particular area to capture any possible changes that may occur within each zone in terms of noise, change of brightness or color, etc. Then the descriptors of each class are analyzed to associate to each region class (j) the weight of intra-class (w_{fj}) respective to each descriptor (f_i).

To determine the weight (w_{fj}), the standard deviation (σ_{fj}) of each descriptor (f_i) is calculated with the class of zone (j). (σ_{fj}) measures the variability of data in a descriptor region class. Thus, the larger the value of standard deviation (σ_{fj}); the more unpredictable the descriptor (f_{i}) is for the region class (j). Then, we calculate the weight (w_{fj}) of the descriptor (f) in the region class (j) according to equation (10).

$$w_{f_j} = \frac{(1 - \frac{\sigma_{f_j}}{\sum_{f=1}^{F} \sigma_{f_j}})}{\sum_{f=1}^{F}(1 - \frac{\sigma_{f_j}}{\sum_{f=1}^{F} \sigma_{f_j}})} \tag{10}$$

Obviously, the weights (w_{fj}) defined by this way for a class of zone (j) must satisfy the following constraint:

$$\sum_{f=1}^{F} w_{f_j} = 1$$

For some samples of TV news selected from different channels, the following results are obtained:

Table 1. Values of weight features of different TV news

	w1	w2	w3	w4
P1	0,331	0,333	0,091	0,244
P2	0,333	0,333	0,016	0,316
P3	0,332	0,332	0,225	0,110
P4	0,332	0,333	0,144	0,190
P5	0,333	0,333	0,008	0,325
P6	0,333	0,333	0,016	0,316
P7	0,333	0,333	0,004	0,333
P8	0,331	0,333	0,200	0,135
P9	0,332	0,333	0,047	0,287

The global zone descriptor for each visual invariant is defined by a weighed summation of these metrics form the texture descriptor (11).

$$\text{desc(Decor)} = w_1 E \cup w_2 LHo \cup w_3 C \cup w_4 Ho \tag{11}$$

5 Grammar of TV News Programs

After having detected and identified the presenter, his descriptors are coupled with the decor descriptors, a single description of a shot presenter is obtained (12).

$$PD_k =< P_k, \{desc(Decor)\} > \tag{12}$$

With each appearance of a new presenter, the process of identification is started in order to validate it. Each new presenter is added afterwards to the grammar's list of presenters already identified in the former iterations. So, there will be ***X*** iterations for identified ***X*** presenters that appear tens even hundreds of times (***N*X***), for only one year of streaming of a channel, resulting in an important reduction of structuring algorithm complexity.

Considering their redundancy, and in addition to the same reasons of complexity reduction, the decor descriptors are extracted just after identifying the first occurrence of the presenters.

Thus, each TV news program is described in the grammar by the descriptors of its jingle for the identification in TV streams and the descriptors of the studio decor in addition to the descriptors of all the presenters who can present this program to be able to deduce his internal structure.

A generalization of the step suggested for the majority of the programs of a channel leads to a grammar/channel composed by all the descriptors of location and the internal structuring of these programs.

The identification and segmentation of the programs of a TV channel are realized through the integration of research and comparing techniques of the descriptors of grammar with those of TV streams.

6 Conclusion

In this paper, the generation of video grammars is approached. The role of these grammars is to model and store a priori knowledge, useful for the structuring of video streams, in the form of descriptors. Indeed, the recurring descriptors are classified into two levels: descriptors for the program identification in streams, and others for the internal structuring of the located programs. For the second type, TV news programs are selected as a case of study. In this type of media, we focus on the presence of the presenters' shots as the internal unit of structuring.

The perspectives are to extend the number of descriptors of intra-segmentation to have almost complete grammars for the various channels. We also think of suggesting and integrating extraction modules of the relevant and common descriptors for the majority of television broadcasts.

References

1. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics 3, 610–621 (1973)
2. Turk, M., Pentland, A.: Eigenfaces for Recognition. Journal of Cognitive Neuroscience 3, 71–86 (1991)
3. Arvis, V., Debain, C., Berducat, M., Benassi, A.: Generalization of the Co-occurrence Matrix for Color Images: Application to Color Texture Classification. Image Anal. Stereol., 63–72 (2004)
4. Zlitni, T., Mahdi, W.: A visual grammar approach for TV program identification. International Journal of Computer and Network Security 2(9), 97–104 (2010)
5. Viola, P., Jones, M.J.: Rapid Object Detection using a Boosted Cascade of Simple Features. In: Conference on Computer Vision and Pattern Recognition, USA, vol. 1, pp. I-511–I-518 (2001)
6. Lienhart, R., Maydt, J.: An Extended Set of Haar-Like Features for Rapid object Detection. In: IEEE International Conference on Image Processing, USA, vol. 1, pp. 900–903 (2002)
7. Jacobs, A., Miene, A., Ioannidis, G.T., Herzog, O.: Automatic shot boundary detection combining color, edge, and motion features of adjacent frames. In: TRECVID Workshop Notebook Papers, pp. 197–206 (2004)
8. Haller, M., Kim, H., Sikora, T.: Audiovisual Anchorperson Detection for Topic-Oriented Navigation in Broadcast News. In: IEEE International Conference on Multimedia and Expo., Canada, pp. 1817–1820 (2006)
9. Zlitni, T., Mahdi, W., Ben-Abdallah, H.: A new approach for TV programs identification based on video grammar. In: 7th International Conference on Advances in Mobile Computing and Multimedia, Malaysia, pp. 316–320 (2009)
10. Misra, H., Hopfgartner, F., Goyal, A., Punitha, P., Jose, J.M.: TV News Story Segmentation Based on Semantic Coherence and Content Similarity. In: Boll, S., Tian, Q., Zhang, L., Zhang, Z., Chen, Y.-P.P. (eds.) MMM 2010. LNCS, vol. 5916, pp. 347–357. Springer, Heidelberg (2010)
11. Zlitni, T., Mahdi, W., Ben-Abdallah, H.: Towards a modeling of video grammar based on a priori knowledge for the optimization of the audiovisual documents structuring. In: 2nd International Conference on Computer Technology and Development, Egypt, pp. 517–521 (2010)

Multispectral Imaging and Digital Restoration for Paintings Documentation

Marco Landi[1] and Giuseppe Maino[1,2]

[1] Faculty of Preservation of the Cultural Heritage, University of Bologna, Ravenna site,
5, via Mariani, Ravenna, Italy
giuseppe.maino@unibo.it,
mrclnd@hotmail.it

[2] ENEA: Italian National Agency for New Technologies, Energy and Sustainable Economic Development, 4, via Martiri di Montesole, Bologna, Italy
giuseppe.maino@enea.it

Abstract. Spectral imaging for radiation wavelengths different from the visible ones, namely in the infrared (IR) and ultraviolet (UV) ranges provides useful information about the actual preservation state and past conditions of paintings. As a consequence, it is possible to combine this information with that obtained in the usual RGB visible basis and to propose digital or 'virtual' restoration of a painting, taking into account its history, modifications and repaintings done in the past. As an example, a work of Pietro Lianori is discussed and analysed.

Keywords: Multispectral analysis, IR and UV images, virtual restoration, painting.

1 Introduction

The virtual restoration is now a good opportunity for analyses to be performed by restorers, conservators and art historians. Many works often can not be restored, sometimes for lack of valid techniques or because of bad previous operations, sometimes due to deficiency of precise records allowing us to understand how the work had been rebuilt.

As an example, when the paintings show successive repaintings, - a well-known case is represented by the painting of Caravaggio, *The good luck*, conserved in Capitolini Museums of Rome – one should not intervene with the removal of the second painting since itself is an artistic work to be preserved, but in digital you have the freedom to remove or divide the two paintings by techniques that allow retouching, with higher magnification, to restore the original features, without the work suffering a loss.

Within a digital framework, there is a large freedom of action to create hypotheses for restoration of paintings, frescoes, but also photographs, architecture and three-dimensional objects. Today, the digital (or virtual) restoration is considered an accepted technique for the restoration of old photographs, but it is also an excellent opportunity to develop possible interventions on paintings, sculptures and archaeological artefacts.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 464–474, 2011.

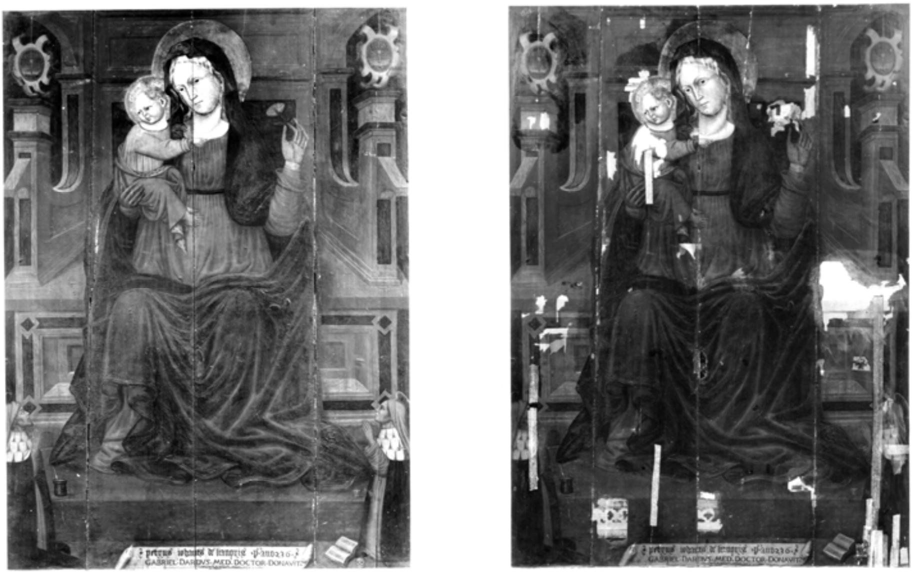


Fig. 1. Pietro Lianori, *Virgin with the Child,* Cappuccinis' Museum, Bologna, XV century, 171 x 124 cm, before restoration (on the left) and during the restoration (on the right)

2 The Virgin with Child of Pietro Lianori

In this paper we consider a very controversial and partial restoration carried out on a painting by Pietro di Giovanni Lianori, representing the Virgin and the Child (fig. 1), an artist active in the fifteenth century. This work is conserved in the Museum of Cappuccini in Bologna, Italy, since 1928. In the provincial archives of the Cappuccini Friars of Bologna the photographs are preserved, documenting the status of the work before and during the restoration that has been interrupted because of the discovery of the original painting below a successive remaking (fig. 2).

These images confirmed a very difficult situation for restorers and conservators, due to a complex overlapping of layers both original and repainted in order to refresh the painting and to highlight the identity of a new donor. In fact, the comparison between data recorded from multispectral images in various spectral bands has allowed the identification of quantitative and qualitative differences subsequent to the drafting of the original (fig. 3).

The comparison of the recovery image of this panel in visible light before, during and after the restoration of the seventies of the twentieth century clearly shows that the work had been almost entirely repainted. If the choice to remove, in some areas, repaintings arising from the seventeenth century modifications made it possible to uncover the original paint surface that is still preserved in the underlying layers, although very impoverished in terms of material, it has also compromised the ability to read all work in a consistent manner.

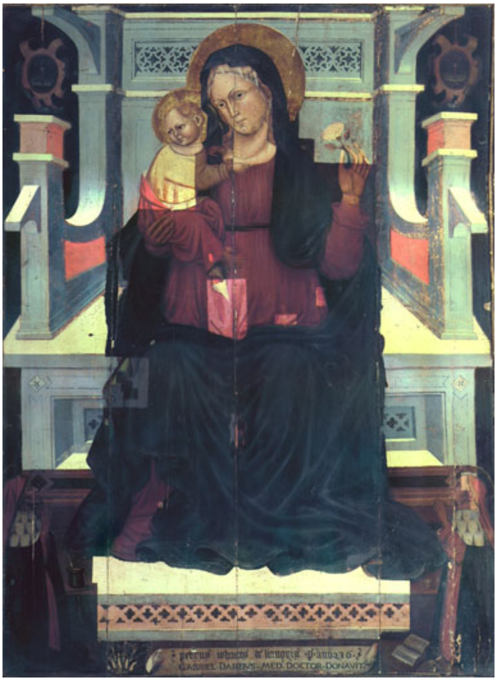


Fig. 2. Pietro Lianori, Virgin with the Child, image in visible light, current issue

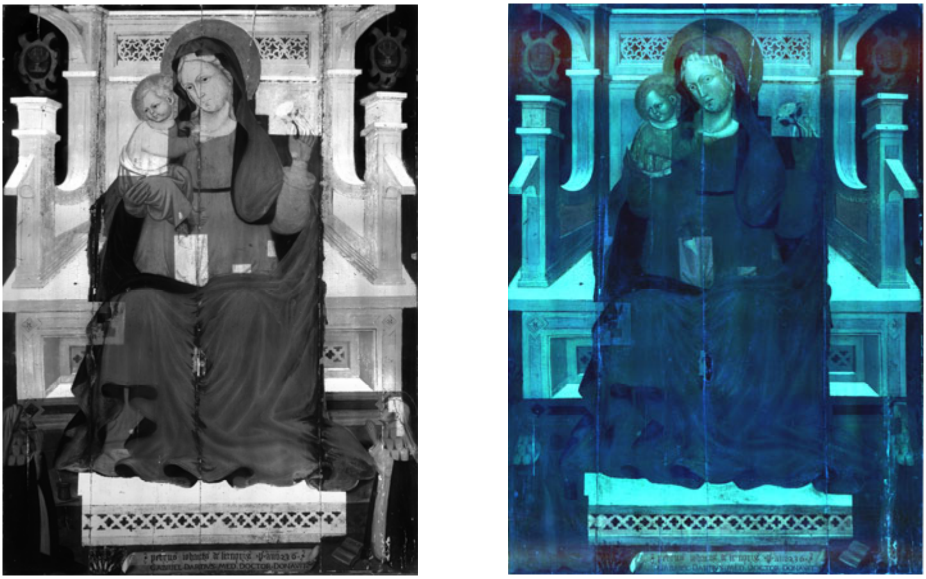


Fig. 3. Pietro Lianori, *Virgin with the Child*, IR image (on the left) and image in UV fluorescence (on the right)

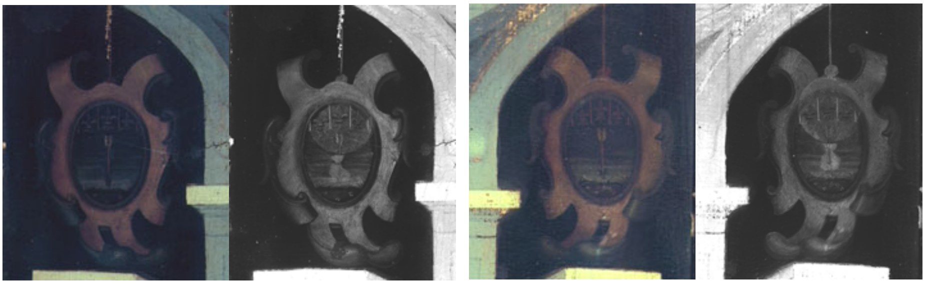


Fig. 4. Pietro Lianori, Virgin with the Child, details showing the two coats of arms in visible light (left) and IR reflectography (right) with different symbols

The dynamic composition of the painting is monumental, a feature typical of the artist, where the Virgin is represented seated on the throne with his right arm raised and holding the Child, and in her left hand there is a white rose, symbol of purity. Above the shelves aside of the throne, one can see two coats of arms that may lead back to bidders, smaller paintings in the lower part of the work, the sides of the central throne (fig. 4).

In the latter there is a scroll with the following inscription:
“PETRUS IOHANIS DE LIANORIS PXT ANO 236 GABRIEL DARDUS MED. DOCTOR DONAVIT ANNO DNI 1611”, written in gothic letters in the first row and in the second one in Roman letters.

By means of multispectral investigation it was possible to examine non-destructively the entire surface of the painting and find some significant areas where the different penetration of the radiation used has highlighted the differences between the drafts and the fifteenth-seventeenth century modified version as in the case of the two coats of arms.

3 Multispectral Imaging and Processing

Multispectral images have been obtained by means of a portable equipment, MUSIS 2007, provided with CCD camera that acquires images with 558 x 370 pixels, recorded at the wavelengths of RGB components of visible light, near-infrared and UV fluorescence. The spatial resolution in the acquisition phase has been chosen in such a way that the digital images could reproduce the *craquelures* of the painted surface; this result has been obtained with a sampling frequency corresponding to 6 pixels/mm.

Therefore, the whole painting has been digitized by capturing partial images referring to an area of 9 x 6 cm and, in order to improve the signal noise ratio, each image consists of an average image of 16 acquired frames for the following recomposition of the whole digital image. The complete final digital images of the painting have been obtained by means of suitable algorithms of reassembly and are shown in figs. 2 and 3.

In order to improve the image contrast, all images have been made uniform radiometrically by means of an equalization procedure and a mean-centering scaling algorithm was adopted [1]. As shown in our previous works [2-5], the multispectral image of figs. 2 and 3 can be summed up in a three-way array, G(IxJxK), with I and J indexes labelling row and column, respectively, of each image, while K refers to the wavelengths. The results of standard PCA (Principal Component Analysis) [6-11] applied to G matrix are summarized by the simple expression:

$$\underline{G} = \sum_{a=1}^{A} T_a * p_a \tag{1}$$

where A is the rank of a two-way G(IxJ)xK matrix resulting from the rearrangement of G, and each term in the summation is multivariate product, T_a*p_a, between the score images, T_a , and the corresponding loading vectors, p_a. Only three PCs explain with “sufficient accuracy” the multivariate image G, and the correspondent score images, T_a, as proved by the structure of the loading vectors. As evident by the loading plots, the first component, T_1, summarizes the information common to all the wavelength with some emphasis on RGB ones. As a consequence, the physical features of the figures and details such as the coats are well outlined. Second and third score images, T_2 e T_3, mainly resume IR and/or UV wavelength concerning the preparatory drawings as well as tonalities present both in the figure of the Virgin and on the background. Moreover, in these score images the restoration areas are clearly identified. T_4 e T_5 images are responsible of a small amount of the whole multispectral variability, whose importance must be evaluated time by time.

This case study shows that the detection of photon spectra combined with the *x-y* scanning procedure is a multivariate measurement. Indeed, *I* scannings of *J* voxels each are performed in order to detect the corresponding scattering spectrum; furthermore, each spectrum is functionally divided in *K* energy intervals, so that a three-way (three-mode) data array follows. In these terms, the layer density evaluation may be considered a spectroscopic imaging technique like satellite and microscopic imaging, where the spatial and spectral data expressed in a *multivariate image* are used to extract significant information [6].

In fig.5A, the three-way data matrix is shown as a stack of congruent images viewing the same field-of-view, measured for a series of different ‘variables’. The *k*= 1, …, *K* frontal slices are two-way images resulting from the integral of the corresponding energy interval computed for all the *I*x*J* voxels. As a consequence, the pixel density distribution comes from the collapse of the multiple images into a single one. The three-way data array shown in fig.5A is denoted by an underlined bold-face matrix, $\underline{\mathbf{G}}^{(\mathrm{IxJxK})}$, and its elements as g_{ijk}, where indices *i*=1, …, *I*, and *j*=1, …, *J*, indicate voxels and *k*=1, …, *K* the energy intervals. Following the terminology used in [9] for the *N*-way image analysis, arrays of three-way data can be characterized by a categorical object/variable (O/V) mode convention, and a multivariate image as an object-object-variable (OOV) array; in these terms voxels in mode A) and B) in fig.5A are the objects, and energy intervals in mode C are variables.The three-way array $\underline{\mathbf{G}}$ may be unfolded into a two-way data matrix and examined with ordinary principal component analysis (PCA). This approach is termed *unfold*-PCA and is summarized in fig.5B, where the stack of images $\underline{\mathbf{G}}^{(\mathrm{IxJxK})}$ is firstly unfolded into a two-way matrix, $\mathbf{G}^{(\mathrm{IJxK})}$, and then represented by means of structure and noise terms.

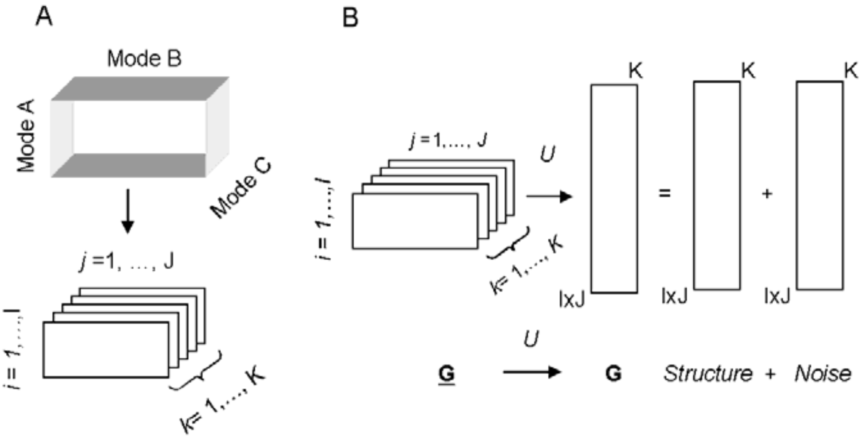


Fig. 5. A) Three-way array, G(IxJxK).. B) Unfold-PCA (see text), U denotes the unfolding operator

The two-way data matrix, **G,** has *IJ* objects corresponding to the spectra detected at each voxel, and *K* homogeneous variables – the number of photons in each energy interval – ranging between tens to tens-of-thousand. In order to take into account the differences between variables, **G** was variable-wise scaled by dividing each vector $\mathbf{g}_k$ by the square root of its mean value, and then centered to obtain a new matrix, **X.** This scaling gives the right weight to every variable in **G** matrix, and makes it possible to obtain the appropriate linear model that describes the data variability.

To reduce matrix **X** into principal components, the singular value decomposition (SVD) was computed,

$$\mathbf{X} = \mathbf{UDV}^T = \sum_{k=1}^{K} d_k \mathbf{u}_k \mathbf{v}_k^T = \sum_{k=1}^{K} \mathbf{t}_k \mathbf{p}_k^T \qquad (2)$$

where vectors $\mathbf{t}_k = d_k\mathbf{u}_k$ and $\mathbf{p}_k = \mathbf{v}_k$ are the scores and loadings vectors, respectively, of the structure term in fig.5B, and K is the full rank of **X** matrix. Scores and loadings vectors in eq.(2) are, respectively, orthogonal and orthonormal.It is found that a two PCs model, accounting for about 92% and 6% of the total variance, respectively, describes with a "sufficient accuracy" the **X** matrix; i.e., the effective rank of **X** is A = 2, and the remaining PCs account for the noise term only of fig.5B. These results were further confirmed by the Bartlett sphericity test with p = 0.9 [10]. The model with two principal components is given by

$$\mathbf{X} = \hat{\mathbf{X}} + \mathbf{E} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \mathbf{E} \qquad (3)$$

where $\hat{\mathbf{X}}$ and **E** matrices are the structure and noise terms, respectively, and the outer products between loadings and score vectors, $\mathbf{t}_k\mathbf{p}_k^T$, univocally identify the structure term.

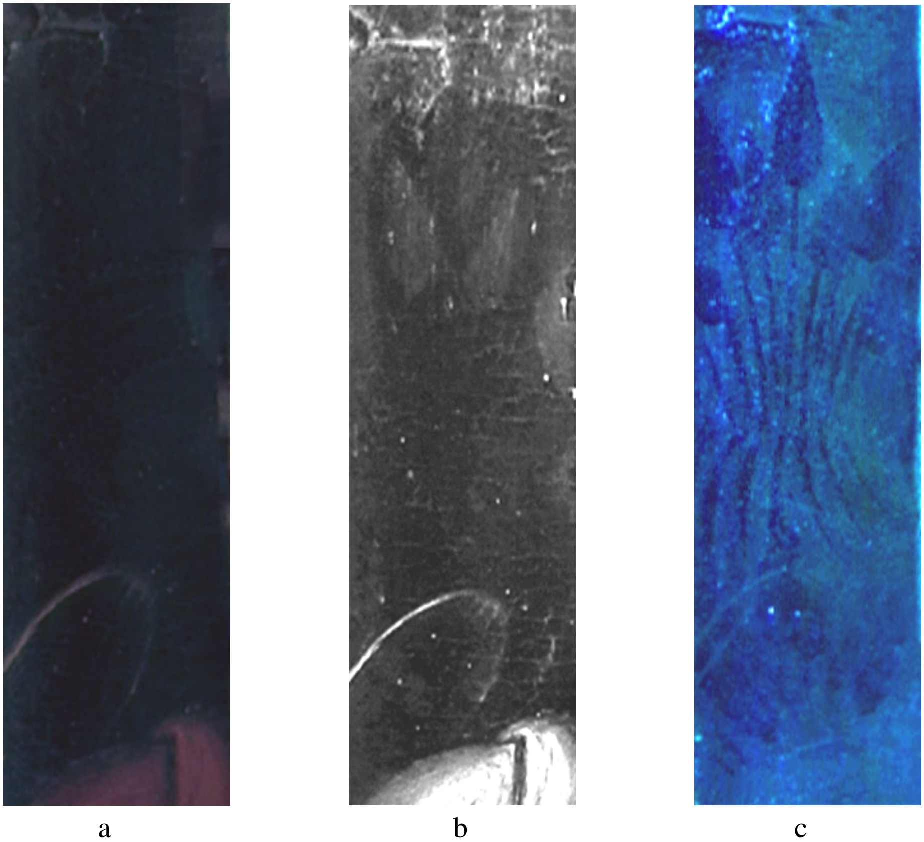


Fig. 6. Pietro Lianori, Virgin with the Child, details in visible light (a), infrared reflectography (b) and ultraviolet fluorescence (c)

Visualization of model parameters, loading and score vectors, has an important role in interpreting the results of multivariate spectroscopy data analysis. Moreover, by the difference between the **X** matrix evaluated from the visible image before the partial restoration and that obtained from the multispectral analysis of the painting in its present status, we derive a quantitative estimate allowing a well-grounded digital restoration to be performed pixel-by-pixel in each region of interest.

4 Results and Conclusions

Figs. 6-8 show a few details of Lianori's painting where digital images have been captured in visible RGB basis, in IR reflectography and UV fluorescence to be combined together according to the procedure summarized in the previous paragraph.

A PCA analysis has been performed for the whole panel, starting from these photon spectra, digitally recorded and processed, as in fig. 5 and eqs.(2,3). Once obtained the **X** matrix a pixel-by-pixel local operations have been carried out by subtracting the content of each pixel in the X matrix corresponding to the RGB image before the restoration.

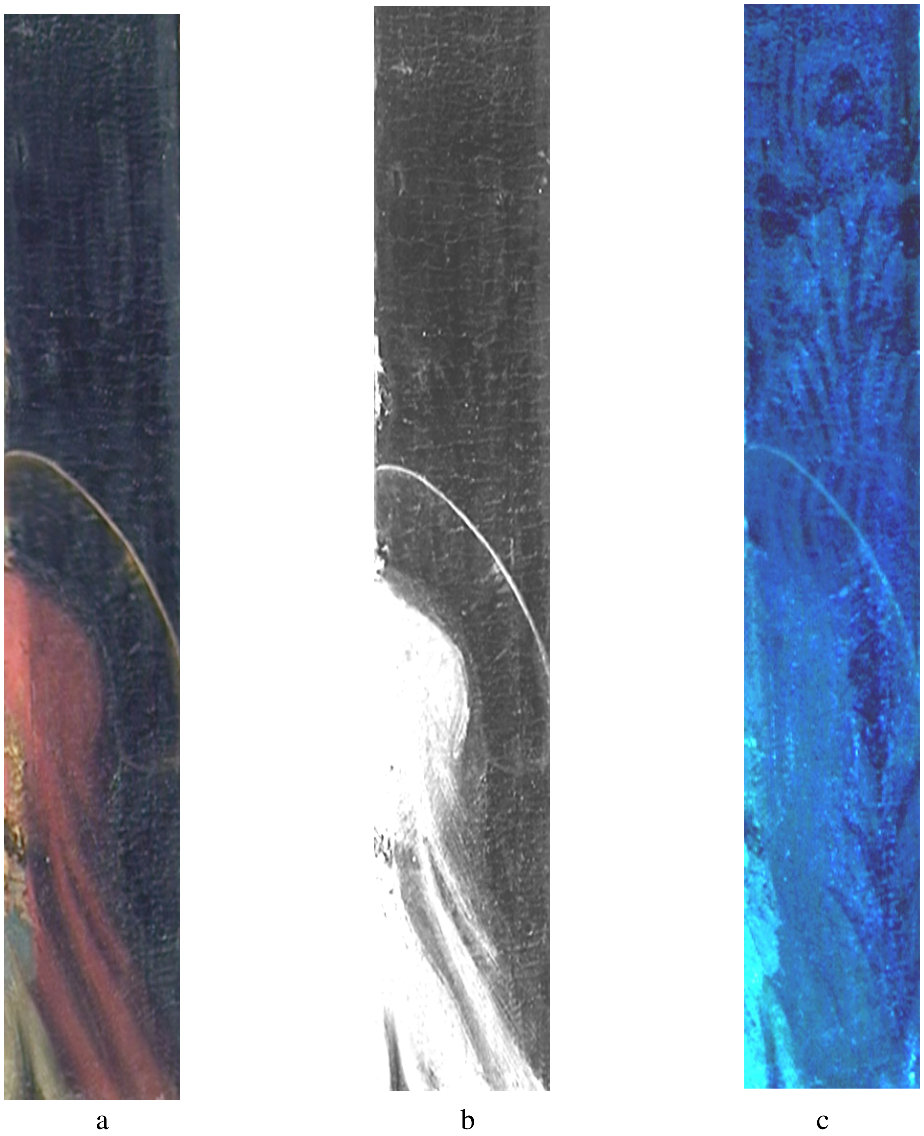


Fig. 7. Pietro Lianori, *Virgin with the Child,* other details in visible light (a), infrared reflectography (b) and ultraviolet fluorescence (c)

Therefore, a digital or virtual restoration of the painting is recovered where the hidden original layer is brought back to pristine condition, as shown in figs. 9 and 10. The concept of 'virtual restoration', despite its introduction is recent enough, is not without ambiguity of use and meaning. Founded in the field of cultural heritage, it has gradually emancipated widening the scope of its application and getting to lick even the linguistic and literary studies, yet the term 'virtual restoration', even in its variant of 'digital restoration', continues to show just the technical means used rather than a clear instance of methodology. Recently, someone has seriously proposed abandonment of these expressions in favor of a more comprehensive definition, but still with a good degree of vagueness, such as, for example, 'iconological digital restoration'.

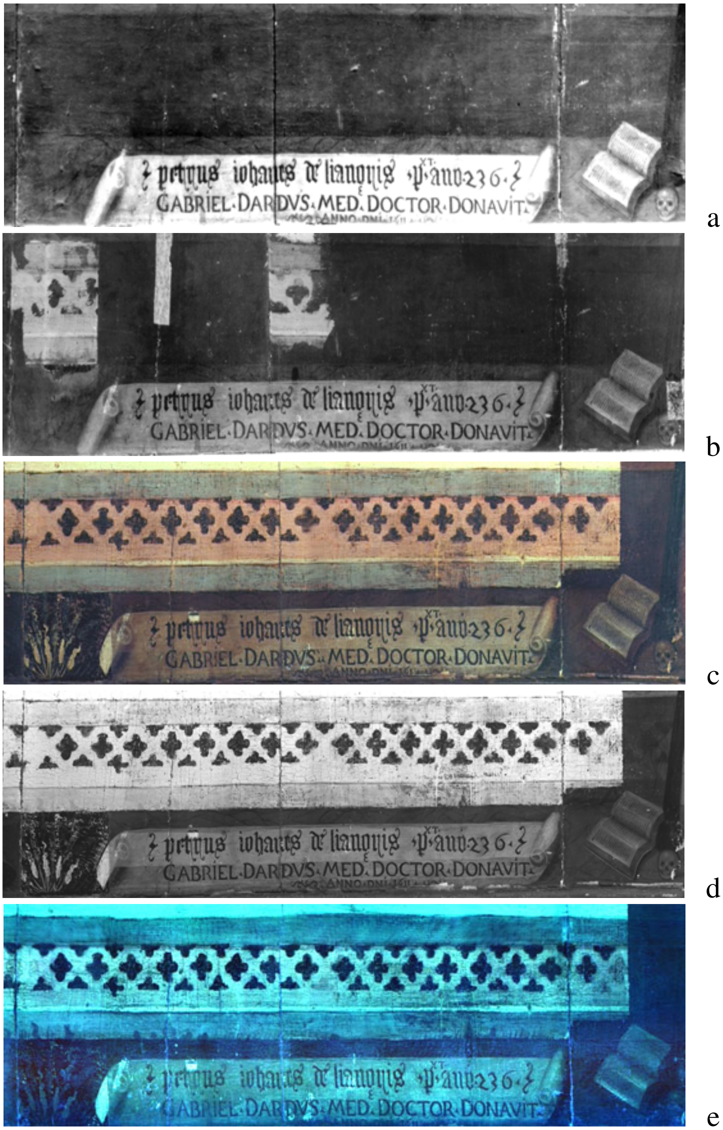


Fig. 8. Pietro Lianori, *Virgin with the Child,* details of the base of the throne with the inscription before restoration (a), during restoration (b), in visible light (c), infrared reflectography (d) and ultraviolet fluorescence (e)

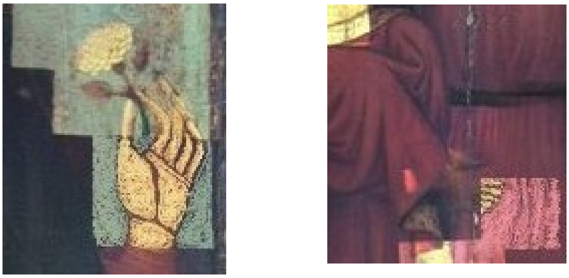


Fig. 9. Two details of Lianori's painting during virtual restoration

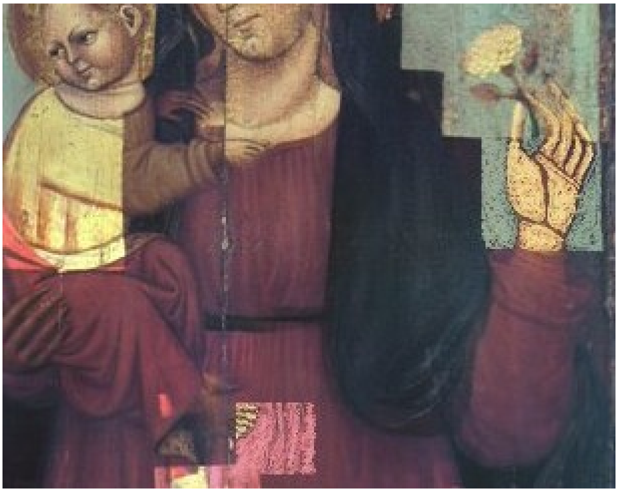


Fig. 10. Preliminary results of virtual restoration

Our purpose was to prove that even a virtual restoration can be performed by respecting the principles stated by Cesare Brandi [13] for real restorations and overall in the world adopted, namely respect for the aesthetic and historical aspects, recognizability of the intervention, reversibility of the materials and minimal intervention, while the fourth one, compatibility of materials, is pertinent only to the real restoration.

References

[1] Gonzales, C., Woods, E.: Digital Image Processing. Addison-Wesley, N.Y (1993)

[2] Bonifazzi, C., Ferriani, S., Maino, G., Tartari, A.: Multispectral examination of paintings and works of art: A principal component analysis approach. In: Proceedings of CLADAG 2003 Meeting of the Classification and Data Analysis Group of the Italian Statistical Society, CLUEB, Bologna, pp. 67–70 (2003)

[3] Bonifazzi, C., Ferriani, S., Romano, A., Maino, G., Tartari, A.: Multispectral Examination of Paintings: A Principal Component Image Analysis Approach. In: Proceedings of ART 2005 – 8th International Conference on Non-Destructive Investigations and Microanalysis for the Diagnostics and Conservation of the Cultural and Environmental Heritage, Lecce, May 15-19 (2005)

[4] Maino, G., Bruni, S., Ferriani, S., Musumeci, A., Visparelli, D.: Multispectral analysis of paintings and wooden sculptures. In: Proceedings of II Congresso Nazionale AIAr Scienza e Beni Culturali, Patron Editore, Bologna, pp. 203–214 (2002)

[5] MUSIS 2007 MultiSpectral Imaging System, Operational Manual, Art Innovation (2007)

[6] Geladi, P., Grahn, H.: Multivariate Image Analysis. Wiley, N.Y (1996)

[7] Baronti, S., et al.: In: Del Bimbo, A. (ed.) Image Analysis and Processing. Springer, Heidelberg (1997)

[8] Geladi, P.: Chemometrics in spectroscopy I: Classical chemometrics. Spectro-chimica Acta B 58, 767–782 (2003)

[9] Huang, J., et al.: Multi-way method in image analysis: Relationships and applications. Chemometrics and Intelligent. Laboratory System 66, 203–252 (2003)

[10] Jackson, J.E.: A User Guide to Principal Components. Wiley Interscience, Hoboken (1992)

[11] Kiers, H.A.L.: Towards a standardized notation and terminology in multiway analysis. Journal of Chemometrics 14, 105–122 (2000)

[12] Biagi Maino, D., Grimaldi, E., Maino, G.: Analisi multispettrali su un dipinto di Pietro Lianori. Archeomatica, 16–20 (2009)

[13] Brandi, C.: Teoria del restauro. Edizioni di Storia e Letteratura, Roma (1963)

Virtual Reality Models for the Preservation of the Unesco Historical and Artistical Heritage

Roberta Menghi[1], Giuseppe Maino[1,2], and Marianna Panebarco[3]

[1] Faculty of Preservation of the Cultural Heritage, University of Bologna, Ravenna site,
5, via Mariani, Ravenna, Italy
giuseppe.maino@unibo.it, Roberta.Menghi@libero.it
[2] ENEA, Italian National Agency for New Technologies,
Energy and Sustainable Economic Development, 4, via Martiri di Montesole, Bologna, Italy
giuseppe.maino@enea.it
[3] Panebarco & C., via Molino, 9, 48121 Ravenna
marianna@panebarco.it

Abstract. Reproduction of space with very high yield of photo-realism, through the use of special three-dimensional modelling techniques, enables the enhancement of the places of the "Great Mtskheta”, the ancient capital city of Georgia, which are mostly in a state of extreme deterioration and abandonment. The production of integrated multimedia content to digital networks, is the ideal solution for a city like Mtskheta: A valid and concrete proposal so as to capture the added value of the opportunities created by the network, making the present reality.

Keywords: UNESCO, tourism, cultural heritage, 3D virtual world, Geographical Information System.

1 Introduction

More than 30 armed conflicts are currently ongoing around the world. Alongside the loss in human lives, more and more valuable heritage sites are turned into battlefield within the war theatres. More historic and archaeological patrimony is being vandalised, looted, illicitly traded spoiling UNESCO World Heritage Listed (WHL) historic city cores. Both the city of Byblos (Lebanon) and the city of˜ Mtskheta (Georgia) - the focus of our project whose preliminary results are discussed in this paper – were just about to be severely affected by war operations respectively in 2006 the first and in 2008 the second. These events posed to the local administrations of both cities an additional concern, alongside that of the safety for the local civil community: Setting protective measures to protect the World Heritage Listed cities from destruction.

Both Lebanon and Georgia are parties of the 1954 Hague Convention (HC). The widespread destructions suffered by ‘cultural properties’ in addition to the huge loss of human lives during World War II, created the ground for the 1954 Hague Convention for the Protection of Cultural Properties during times of armed conflicts to be issued and ratified alongside its first protocol.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 475–485, 2011.

The 1999 Second Protocol of the Convention entered into force in 2004 and in 2008 a Committee set at UNESCO finally issued the Guidelines for the implementation of the Convention. Over 56 years from its first ratification the HC remains widely neglected due to a lack of instruments for its implementation. In 2009 Guidelines for the implementation of the Convention and its two 1954 & 1999 Protocols were issued and now we have conceived and developed a WAR FREE WORLD HERITAGE LISTED CITIES (WFWHLC) project to set a very concrete model in Byblos and Mtskheta, two cities that are patrimony of the humanity which were recently threatened by war.

This WFWHLC project is funded by the European Union within the framework of the CIUDAD program for a duration of 30 months and a budget of 540,740 Euro. Main project locations are Lebanon and Georgia and participants are the local governments of Jbail-Lebanon and Mtskheta-Georgia, WATCH – an NGO linked to UNESCO - in association with FOCUH (Turkey) and NEREA (Italy). NEREA is a joint venture of the University of Bologna and ENEA, Italian National agency for new technologies, energy and sustainable economic development. A first kick off meeting among the partners took place in Byblos, October 28-30, 2010.

The 1954 Hague Convention provides for the protection of 'cultural property' three moments, namely before, during and after armed conflicts; they coincide with three different approaches to be undertaken by national and/or international Cultural Heritage operators deployed in areas of conflict:

- Risk preparedness (and mitigation).
- Risk Management.
- Post conflict operations.

lt would be obvious that protective measures for the two cities should be inspired to the HC. However, one of the main contingencies faced by the concerned authorities in the implementation of the Convention and its Protocol is to be associated to the very limited availability of concrete case study that could serve as a reference when formulating their Risk Preparedness Plans (RPP) or when applying to UNESCO - Committee for the Protection of Cultural Property in the event of armed conflict (Committee) for the introduction of protective measures. The subject requires a comprehensive approach as well as a solid policy for the maintenance of the protective measures during times of conflict. This implies the definition of a well designed pattern of relations between Civil and Military Authorities as well as the Civil Society at large and, first of all, the implementation of a multimedia database where all the information referring to the considered cities have to be stored and processed.

Based on these considerations WFWHLC intends to contribute to the definition of a comprehensive urban planning strategy that is designed to introduce and maintain the Enhanced Protection status for WHL cities. Therefore, overall objectives of our project are the development and applications of computer science tools in image and data processing in order (i) to create safety conditions for two WHL sites under immediate threat, (ii) to set good practice in the Urban Planning/Management of WHL cities under latent threat of armed conflicts, and (iii) to promote a widespread awareness about risks and risk mitigation measures needed to secure Enhanced

Protection for WHL cities threatened by aimed conflicts, in accordance with international conventions.

Specific objective is to produce a Risk Preparation Plan and promote awareness about threats posed from armed conflicts for two WHL cities in support of the concerned authorities to candidate Byblos and Mtskheta for Enhanced Protection from UNESCO. Apart from the evident immediate effects produced by the project for the protection of Byblos and Mtskheta, the achievement of the project objectives will contribute to set several milestones, the most important of them being the methodological approach that will be disseminated to help other WHL cities that are experiencing the risks of conflicts and/or natural disasters without proper risk preparation and preventive plans.

Finally, major expected results include risk plans for the two WHL cities, risk mitigation policies in place, main actors, stakeholders and civil society awareness promoted by setting practices for the implementation of The Hague Convention (1954) Guidelines. The main tool to achieve these results is the implementation of a multimedia database where an open source geographical information system (GIS) is coupled to a digital platform for 3D interactive simulation of real world, freely available to the concerned and interested people. A preliminary analysis has been performed on Mtskheta city and is presented in this work.

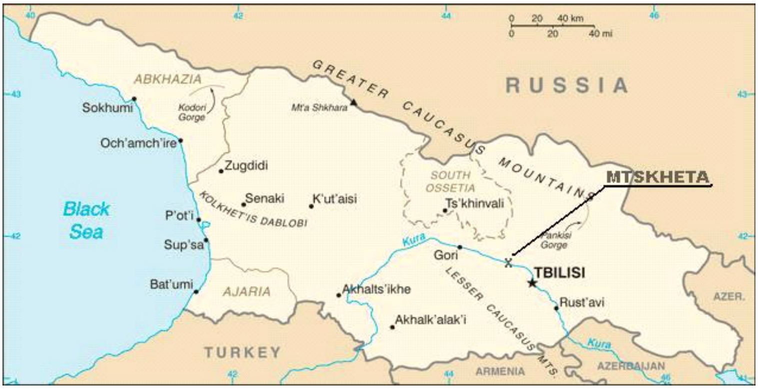


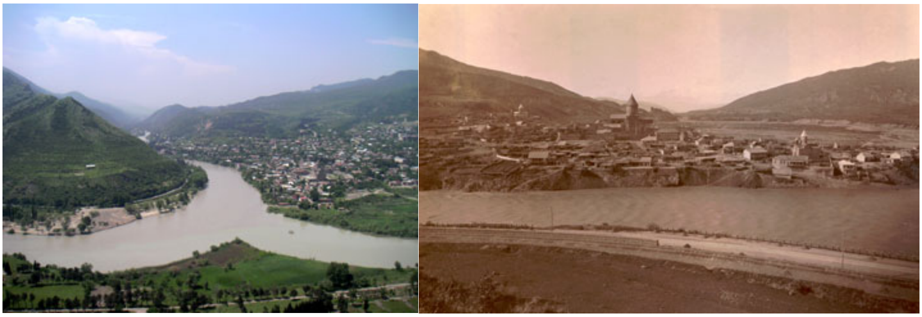
Fig. 1. Location of the city of Mtskheta

2 Mtskheta Short History and Cultural Heritage

In 1994 Mtskheta has joined the UNESCO World Heritage List for its historic medieval churches risen to one of the most significant examples of the early Christian architecture in Georgia, with a strong impact on the subsequent development of Georgian architecture. The document of the 18th session of the World Heritage held on 12 to 17 December 1994 in Phuket, Thailand states:

<<The Committee, in inscribing this property on the World Heritage List, suggested to the State Party to change the name to "Historic Churches of Mtskheta (…) That this property be inscribed on the World Heritage List on the basis of criteria

iii and iv: Criterion iii - the group of churches at Mtskheta bears testimony to the high level of art and culture of the vanished Kingdom of Georgia, which played an outstanding role in the medieval history of this region; Criterion iv - the historic churches of Mtskheta are outstanding examples of medieval ecclesiastical architecture in the Caucasus region>>.



Figs. 2. & 3. Different views of Mtskheta

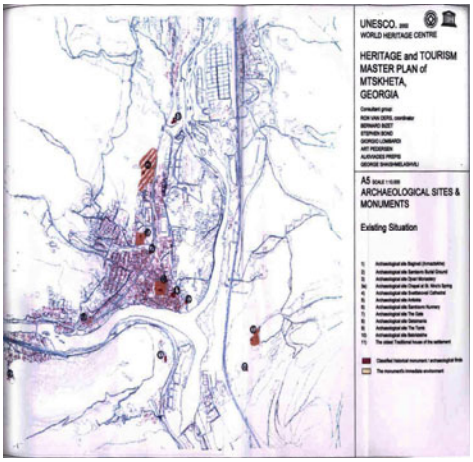

UNESCO
WORLD HERITAGE CENTRE
HERITAGE and TOURISM MASTER PLAN of MTSKHETA, GEORGIA
A5
ARCHAEOLOGICAL SITES & MONUMENTS
Existing Situation

Fig. 4. Map of archaeological sites and monuments of Mtskheta (1:10,000), included in UNESCO World Heritage List in 1994

In 2005, the name of the property was changed and it was called the "Historical Monuments of Mtskheta". In 2009 it was included in the UNESCO list of Heritage in Danger, with the requirement from the state party to specify which monuments have been placed on the list. Actually, the National Agency for Cultural Heritage Preservation of Georgia works jointly with the UNESCO and re-nomination of monuments of Mtskheta is a possibility to address the problem.

In August 2008, the Russian army moved towards Tbilisi and reached environs of Mtskheta. Russian troops were in few kilometers from the city when the war was stopped due to the mediation by the President-in-Office of the European Union, French president Nicolas Sarkozy. The town of Mtskheta and its historical monuments remained intact during the conflict. However, a number of monuments in and around Tskhinvali region/South Ossetia were heavily damaged during the Russian-Georgian war.

Mtskheta is the ancient capital of Georgia, nestled in a picturesque setting located in a valley at the confluence of the rivers and Aragvi Mtkvari (figs.1-3) and is 20 / 25 km north-west from Tbilisi (now the capital of Georgia). From a moral point of view, Mtskheta has a formidable impact on the Georgians, as Masada was for the ancient Jews. The site is in fact a kind of inner sanctuary of artistic and religious culture of the country: Here, in 334 AD Georgians were converted to Christianity (according to history they were the second people to take the plunge, just four years after the neighboring Armenians) and Mtskheta, from III BC to V AD (see the archaeological sites depicted in fig.4) was the capital of the Kingdom of Georgia (even after the capital in the sixth century became Tbilisi, Mtskheta continued to be place of coronation and burial of the Georgian kings until the end of the Kingdom that took place in the nineteenth century). Christianity was then introduced in the fourth century in Mtskheta, and more generally in Georgia, by Saint Nino, and the first wooden church was built in the garden of the Royal Palace where now the cathedral of Svetitskhoveli stands (fig.5).

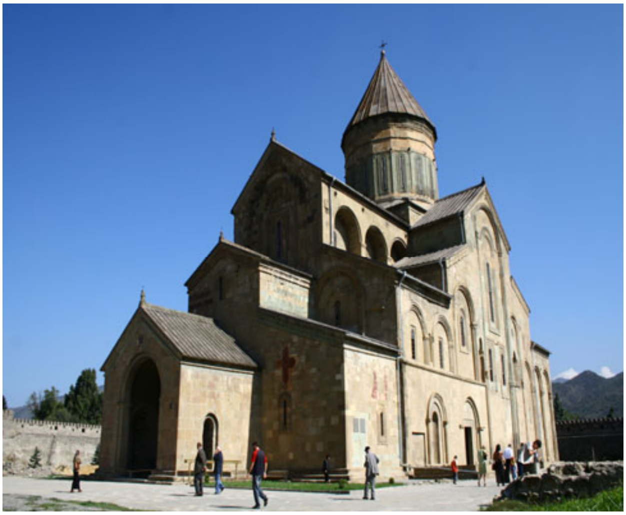


Fig. 5. The cathedral of Svetitskhoveli

Svetitskhoveli Cathedral (XI century) and Djvary Monastery (IV century) are still major landmarks, symbols of the country, together with the fortress of the Acropolis of Bebris Tsikhe, built by Armaz Tsikhe in the III century. The city is still the seat of the autocephalous Georgian Orthodox Apostolic Church: It has therefore retained a central role as a religious center of the country and home to the Katolicos (later elevated to Patriarch), an important reference point for the Georgians and the believers in other parts of the world.

Despite the cultural importance and the many churches (essential for the development of medieval architecture in the Caucasus), the modern city in appearance is rather modest, both in terms of the size and architecture of today's buildings. An added value of the place lies in its natural arena, the environment particularly impressive that can be viewed from various sites, providing a smooth, pristine landscape that contributes to the small size of the city and to its architecture that blend in harmony with nature.

Moreover, because of the serious condition and, in some cases (e.g., the Getsimania church), abandonment of some monuments and recent improvident works of 'restoration' carried out which have partially eliminated the historic features of the monuments, most of the archaeological sites has not any form of protection, simply a few ways to shed and fence protection. In the case of the Necropolis of Samtvro which has a simple roof to repair a large area of great historical and artistic value from the weather, existing facilities are clearly insufficient to protect the sites from the natural elements, but also from possible theft or damage caused by man.

A significant example of wrong intervention according to the UNESCO Commission is given by the present situation of the above-mentioned Svetitskhoveli Cathedral, where the ecclesiastical authorities have carried out irreversible restorations, by the use of reinforced concrete and by removing important ornaments of the cathedral, with no previous agreement with local authorities and violating the International Convention for the protection of UNESCO:

<<(…) despite its great importance, it is now in imminent danger due to the large-scale interventions of the local church authorities. These interventions, which have been carried out on the authorities' own initiative in the name of the monastery's current operational needs without any control by the local or central services responsible, are beyond any scientific ethics concerning protection of and respect for monuments and they are in blatant violation of UNESCO's international protection convention>>

The Committee of UNESCO has formally asked Georgia to start a serious work of recovery and maintenance, given the alarming situation in terms of the various historic buildings:

<<(…) the state of conservation of the archeological components of the World heritage property, their progressive deterioration and the abandonment of the conservation efforts by the state party, nothing that this loss has a major impact on the outstanding universal value, authenticity and the integrity of the property and further urges the state party to develop a special program on protection of all archeological components>>.

In order to "remove" Mtskheta and its monuments from the list of monuments in danger, a fundamental aspect is the (reborn) awareness of local political authorities with regard to maintenance necessary for the preservation of the heritage of the city, although there are difficulties in involving all stakeholders in the management. <<The establishment of the UNESCO and International Relations Unit at the National Agency for Cultural Heritage Preservation and the establishment of the State Commission for World Heritage are indicators to the achieved progress>>.

In this respect, GIS archives and 3D simulations of virtual reality may represent a valuable help for local authorities, restorers and scholars like art historians and archaeologists to perform these duties.

3 Mtskheta Virtual World

What is 3D? Essentially, a form of representation of space in its three dimensions on a two-dimensional plane. The 3D technology is expanding for years with a trend of exponential growth. An increasing number of sectors that use the techniques of three-dimensional reconstruction: From aeronautics to medicine, from biology to chemistry, from cinema to television.

In the cultural heritage sector it has been widely used for 3D reconstruction of places that no longer exist or have been partially destroyed; many are the examples of reconstructions of archaeological sites, thanks to the guidance of scholars and art historians. But the 3D can also be used as a tool to conserve, preserve and promote cultural heritage. More and more repositories in the future will spread out in three-dimensional objects, buildings, monuments and archaeological sites will be associated not only to descriptive text files and images but also to detailed three-dimensional documents available at several levels.

The use of 3D will represent a rich and exciting development of archives: The cataloguing of objects, buildings and monuments in the future will necessarily include, in addition to traditional data and details (see cards for works of art) also a further field that will contain a link to the three-dimensional reconstruction, useful for the scholar to the architect, for the student to the restorer, for the curious ...

A virtual world is a place simulated by computer graphics, where users - connected to the Internet - interact via avatars. The term comes from Sanskrit and Hindu mythology and religion, and means taking a physical body by God. An avatar is a digital representation of the user who enters a virtual world. The most popular virtual worlds include MMORPGs (Massively Multiplayer Online Role-Playing Game, online games that simultaneously connect thousands of people around the world). What are the main features of a virtual world?

Sharing experiences: Multiple users can, everyone at her/his desk, log on simultaneously in the world;

real time: The user decides to "live" what to do, how to get around, where to go and everything happens at the very moment in which the actions are carried;

interactive and participatory creation: The user can click on objects, modify them, build new ones;

socialization: Users can socialize and create groups and communities through various communication channels;

persistence: The world exists regardless of whether users are connected;

involvement: The three-dimensionality and esplorabilità the digital space, combined with interactivity and allow a greater degree of social interaction and involvement.

We present a brief description of VirtualLife, an experimental platform whose development is currently in progress, and is the subject of a project co-financed by the European Union under the Seventh Framework Programme – ICT. The aim of the consortium that develops VirtualLife (9 partners from 9 different EU countries, including 7 universities and 2 SME) is the creation of a virtual world platform innovative in terms of both technical and philosophical items, that solves some of the main limitations of existing parallel universes. The research teams of VirtualLife are focusing more on administrative and legislative aspects and on security and data protection.

Basically the project aims to create an immersive virtual environment based on a peer-to-peer network (which technically provides benefits primarily on the distribution of computational load) and where the many capabilities offered by the programming language are combined with the reliability of a secure communications infrastructure, dedicated not only to entertainment but also to socialization, training and business.

One of the main aims of the project is thus to create, in a collaborative way, the first Virtual Constitution, for the creation and administration of the new virtual nation that will come to be absolutely necessary to safeguard the common interests of citizens and private virtual visitors. This opportunity opens the way for the creation of immersive visits in real time of the various archaeological sites scattered in the valley where it stands Mtskheta, visits that can be used by remote users connected to Internet broadband. The project is targeted to the global market, which intercepts the growing demand of the people's network of rich, charming and engaging experiences, so extending the visibility of Mtskheta and its heritage.

Working in this direction will seize a number of advantages:

The overall visibility and global, with the 3D modeling of heritage and its dissemination on line, will help to increase the knowledge of monuments, places and archaeological sites that are now neglected, in a manner that seriously compromises their survival (especially those outside the perimeter of the city). Only through the dissemination of knowledge we can preserve our heritage.

Formation of a three-dimensional archive, the creation of a database of three-dimensional digitized documents, represents for Mtskheta (but not only for this town) the opportunity to recreate the perception of the environment, simulating the look of the place of the old "residents." This is a resource, potentially limitless, which remains "forever" to a past, whose existence will be passed on to future generations.

New perspectives of knowledge, using a multi-dimensional dynamic management for Mtskheta and its heritage, can be developed for each archaeological site, an apparatus of multimedia (eg, cards, audio, video) and information data to emphasize or devise routes that provide movements and points of view impossible in the real world, so as to return to each of the items emphasizing in particular interesting anecdotes and stories, explaining the context in which the work was carried out (interestingly, in these cases, it is also the approach to virtual restoration). A captivating experience that can arouse emotions in the user, such as a visit by the "fly" to match the decor (for example, the frescoes in the dome of the cathedral, the complex of Djvari) in real time and absolute freedom. By this way the visitor has a virtual unique and unforgettable experience. All this information is also useful for those planning a trip to Georgia and also for those who, once back home, will reinforce what they admired in the just concluded trip, taking advantage of opportunities that in the physical world would be a complex mechanism for cost and logistical problems.

Economic benefits, the development of multi-dimensional, against an initial investment due to the cost of modeling only allows the worldwide visibility, greater knowledge of the city, thus possibly leading to an increase in tourism and foreign investment: Revenues may be, in part, provided to the agencies in the restoration and establishment of appropriate structures for the protection of archaeological sites located on the whole area.

It should be emphasized that an experience like this, and more generally the application of virtual reality methods to heritage, in this case the city of Mtskheta, must not be regarded as a debasement of the real assets, but rather an added value and valuable incentives for the conservation, preserving the existence of places, buildings and other property protected by a disclosure difficult to achieve with the "traditional" media (especially in those areas, such as Georgia, suffering from general security problems, due to the turbulent past / present). Just the transformation of the historic and artistic heritage of Mtskheta, opened to the world, cosmopolitan, through virtual tours, would be able to promote a collective cultural growth, as a guarantee "for the safety" of the precious heritage preserved here, property of humanity, in fact.

In the specific case of Mtskheta it could be envisaged to achieve the following 3D locations: Hypothetical reconstructions of sites partially destroyed due to severe neglect and the "turbulent" past of the city of Mtskheta. Some examples include the site of the Valley Armatsikhe, the acropolis of Mtskheta once ancient royal residence for centuries, thus placing virtual sites of origin of objects that are currently in the State Museum of Janashia and in the ethnography and archeology museums of Mtskheta, as well as all the archaeological finds discovered in the tombs of the Necropolis of Samtavro and Armatiskhe, the area where once stood the residence of King Pitiakhshi.

Other actions include possible reconstructions of finds missing or partially destroyed, the ability to create virtually murals, frescoes, paintings that have disappeared with the passage of time or whose conditions are so degraded as not to allow the vision; a significant example is in the village of Dzalisi, not far from Mtskheta, the mosaic floor, known as "The House of Dionysos", dating from the third century. The Svetitskhoveli cathedral has undergone several restorations over the centuries inside, that have eliminated many of the frescoes by altering the original appearance. There the mode of "flight" can be used in real time and absolute freedom; Svetitskhoveli cathedral is richly painted inside with exceptional frescoes dating from the seventeenth century, depicting scenes from the Gospel and the Old Testament, episodes of the Conference of Kartli, portraits of historical figures.

Reconstructions are made possible of the various periods of development of a complex site, as the Monastery of Djvary that includes several buildings constructed at different times, moreover giving the user the ability to move easily from one level to another time and to compare the configurations of the site at various times.

Objectives are the overall visibility and global formation of a three-dimensional archive; new perspectives of knowledge; economic benefits. The project aims to achieve a broad consensus from potential users such as different tourist segments; circuit of the school network; professionals.

4 Vectorial GIS Supporting 3D Virtual Reality Models

Quantum GIS (QGIS) is a user friendly Open Source Geographic Information System (GIS) licensed under the GNU General Public License. QGIS is an official project of the Open Source Geospatial Foundation (OSGeo). It runs on Linux, Unix, Mac OSX, and Windows and supports numerous vector, raster, and database formats and functionalities. Quantum GIS provides a continously growing number of capabilities defined by core functions and plugins. You can visualize, manage, edit, analyse data, and compose printable maps. The major features include:

View and overlay vector and raster data in different formats and projections without conversion to an internal or common format;

create maps and interactively explore spatial data with a friendly graphical user interface;

create, edit and export spatial data using digitizing tools for GRASS and shapefile formats, the georeferencer plugin, GPS tools to import and export GPX format, convert other GPS formats to GPX, or down/upload directly to a GPS unit;

perform spatial analysis using the fTools plugin for Shapefiles or the integrated GRASS plugin, including map algebra, terrain analysis, hydrologic modeling, network analysis, and many others;

publish maps on the internet using the export to Mapfile capability.

In addition, our research group utilizes GRASS GIS (Geographic Resources Analysis Support System) that is an open source, free software Geographical Information System (GIS) with raster, topological vector, image processing, and graphics production functionality operating on various platforms through a graphical user interface and shell in X-Windows. It is released under GNU General Public License (GPL). GRASS contains over 350 programs and tools to render maps and images on monitor and paper; manipulate raster, vector, and sites data, process multi spectral image data and create, manage, and store spatial data.

Originally developed by the U.S. Army Construction Engineering Research Laboratories (USA-CERL, 1982-1995), a branch of the US Army Corp of Engineers, as a tool for land management and environmental planning by the military, GRASS has evolved into a powerful utility with a wide range of applications in many different areas of scientific research.

5 Conclusions and Perspective Work

The three-dimensional development of the cultural heritage by creating a parallel world of multi-dimensional goods integrated to telematic networks does not want to replace the physical model, but rather making it possible to support it with the provision of services otherwise too expensive or even impossible.

The places, squares, monuments, museums, mansions, palaces, churches and all the excellence of the artistic and historical heritage, are to most people silent witnesses of their own history. Through their three-dimensional processing, digital clones, it is technically possible let them interact, through the development of open and dynamic content that will deepen their story, giving the users the feeling of a "live" presence. In addition to the perspectives outlined above, it is interesting to mention the development of other applications, which are present not only *in situ* and remotely accessible. We refer to the installation of multimedia interactive stations (totems), scattered within archaeological sites. These are very powerful communication tools, dynamic and flexible enough to guide the visitor on a tour of "increased" reality and full of content (with the possibility of hypothetical reconstructions of the site, etc..), through the use of various interactive techniques such as video, keyboard and touch screen.

The use of guides on handheld devices, mobile guides, are able to deepen and complete the travel experience of the user, providing the opportunity to take guided tours or access to detailed information concerning the site or relating to specific items on display (see the many, and unknown graves in the Necropolis of Samtavro). The use of mobile guides would also be useful to the visitor during the trip, providing useful information about routes and directions to take (given also the scarcity of signs throughout the area of Mtskheta). Also it is worth mentioning a final interesting aspect of mobile guides, which deal with multi-users to promote social interaction among visitors: The multi-user visits and collaborative group is a key to a successful learning environment.

The Network and the three-dimensional shapes of the real world where you can create new scenarios assume the building of parallel worlds free from physical laws, against initial investment and operating costs very low. Far from thinking of the latest technology as a substitute, it means a direct and personal dialogue with artistic and historical assets. We turn to this as a means of knowledge aimed at overcoming the limits of sense perception, identifying possible solutions to be adopted for the radical change in information technology, without having to invest too many energy, resources, time, learning tools, which require skills of other professional profiles>.

The methods of virtual reality considered in this work to take advantage of multimedia for planning and creating a rich calendar of events parallel to the normal activities in the physical world, easily fit into a broader discussion of preservation and enhancement, decline in contexts where the infinite dimension "virtual" protect and preserve the assets which are also at risk (in most cases due to human intervention).

References

UNESCO, 18com XI-inscription: The City-Museum Reserve of Mtskheta, Georgia (1994), http://whc.unesco.org/en/decisions/3217

UNESCO, 33COM 8C.1 - Update of the List of World Heritage in Danger (2009), http://whc.unesco.org/en/decisions/1983

VirtualLife (2010), http://www.ict-virtuallife.eu

KUNSTHISTORICHES INSTITUT in FLORENZ, Church of Jvari, Cathedral of Svetitkhiveli,

http://expo.khi.fi.it/galleria/georgia/katli/chiesa-di-javari-mtskheta

http://expo.khi.fi.it/galleria/georgia/katli/cattedrale-di-svetitskhoveli-mtskheta

QGIS: http://www.qgis.org

GRASS: http://grass.bologna.enea.it

Image Processing and a Virtual Restoration Hypothesis for Mosaics and Their Cartoons

Mariapaola Monti[1] and Giuseppe Maino[1,2]

[1] Faculty of Preservation of the Cultural Heritage, University of Bologna,
Ravenna site 5, via Mariani, Ravenna, Italy
mariapaola.monti@studio.unibo.it, giuseppe.maino@unibo.it,
[2] ENEA, Italian National Agency for New Technologies, Energy and Sustainable Economic Development, 4, via Martiri di Montesole, Bologna, Italy
giuseppe.maino@enea.it

Abstract. We present some results of image processing relevant to mosaics. In particular, the three-dimensional image of a mosaic is obtained by means of a laser scanning technique, then adapted to a GIS software, describing main characteristics of single tesserae. Furthermore, the virtual restoration of Ravenna mosaic cartons is performed and shown by a few examples.

Keywords: Mosaics, laser scanning, Geographical Information System, virtual restoration, cartoon.

1 Introduction

Cesare Brandi in his *Theory of Restoration* (1963) [1] stated that "restoration is the methodological moment of recognizing a work of art in its physical consistency and in its dual aesthetical and historical polarity, in view of its transmission in the future", that "it must aim at restoring the potential unity of the work of art, provided this is possible without making an artistic or historical fake, and without erasing any trace of the passing of the artwork through time", that "we can restore only the material of the work of art" and that "the restoration, in order to be a legitimate operation, should not assume neither the time to be reversible nor the abolition of history".

Cesare Brandi developed his theory of restoration over the years (1939-59), when he directed the Istituto Centrale del Restauro (ICR), founded in Rome in 1939 by himself, designed by Giulio Carlo Argan, in order to set the task of restoration on a scientific basis and on a multidisciplinary approach, which involves the collaboration of art historians, restorers, and laboratory technicians.

These principles, as intended by the author, clearly referred not to the virtual restoration but to the real restoration, which concerns the artwork substance. Today they are considered as fundamental principles in any traditional restoration intervention. These principles were developed between 1939 and 1959, when virtual restoration did not exist, as we can begin to talk about digital era only since the 60s of XX century [2].

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 486–495, 2011.

However, they - summarized as respect for the aesthetic and historical aspects, compatibility of materials, recognizability of the intervention, reversibility of the materials and minimal intervention - must also be applied to virtual restoration (which must be a legitimate intervention and not a work of fiction or a display of skill), although this consists only in a restoration of an aesthetic nature, which does not involve the material of the work of art. For this reason, the principle of compatibility of materials is not applicable, but all the others remain valid.

Therefore the restoration techniques used via computer, initially called “Electronic Restoration”, now commonly called “Virtual Restoration”, would be more correctly called “Digital Iconological Restoration” because it is a digital processing of the computer image (in Greek είκών) of a work of art [3]. Generally this digital processing is designed to improve the visual and aesthetic aspects of an artwork or to a hypothetical reconstruction of it, which allows for greater readability.

In this sense, the virtual restoration does not compete with the real one but it supports and assists, as it envisages the possible outcomes (such as when you must choose between different methods of intervention) and observes the same rules (arbitrary rebuilding or reconstruction are not permitted). It also allows you to obtain a usable image of the work of art, when the same is not materially restorable because of high costs or a condition of excessive fragility of the materials.

Simplifying, we can say that - while the traditional restoration aims primarily at extending the life of the product -, the virtual restoration aims at achieving better readability of the information contained in the artwork.

We must never forget that a restoration intervention, as in itself always traumatic for the piece of art, should be carried out only in case it is necessary for the survival of the artwork; therefore, when it is necessary to improve the readability of the image, then traditional restoration can be replaced by virtual restoration, thereby conserving the integrity of the original materials. In other words, you can place side by side the fragmented piece and its reconstructed image, cleaned up or rebuilt, but only where it can be done without falling into arbitrary solutions.

Today, “virtual restoration” also means digital improvement, thanks to special algorithms and mathematical relationships, of the images obtained by diagnostic investigations on works of art and antique documents (photographic analysis, X-ray, UV, IR tests), in order to facilitate their comprehension [4]. The advantages of a virtual restoration rely on its specificity compared to the manual method: A digital image, “clone” of the real one, may be altered, duplicated, restored many times without jeopardizing or damaging the real work of art. So, you can work with the maximum freedom of action, in some cases even putting aside the principles of traditional restoration.

Intervention attempts may be different and may also be modified subsequently: Each phase of the intervention may be registered at a different level in a photo editing program like Adobe Photoshop, Corel PhotoPaint or GIMP. In a certain sense this allows for the reversibility of the intervention and immediate comparisons between the different phases and possibly among several operational choices.

The reconstruction hypothesis digitally carried out may also be mimetic, maybe carried out on a different level compared to the original image in order to guarantee their identification. It is important, though, that they be justified by a philological analysis of the piece.

2 Three-Dimensional Laser Scanning of Mosaics

Mosaics have an intrinsic three-dimensional structure due to finite size, shape, position and orientation of tesserae in order to produce particular effects by light reflection, formation of shadows, etc, as well as an extrinsic one since they are often located on curve surfaces such as vaults, domes, pillars and so on. Unfortunately, the usual photographic documentation does not account for these peculiar characteristics and propose a necessarily planar image of mosaics, thus neglecting important information that, in the case of surveys preliminary to restorations, for instance, is rendered in graphics by means of conventional notations.

Therefore, use of 3D laser scanner has been proposed to overcome these difficulties and provides archaeologists, art historians and museums keepers with suitable tools for a better knowledge and representation of mosaics. The instrumentation described in the previous sections has been utilized on samples and large mosaics such as those in the Basilica of St.Apollinare Nuovo in Ravenna (Italy).

A main difficulty arises from lack of data in some regions in correspondence with tesserae borders and dark colour elements – in particular for black tesserae where the laser light is completely absorbed and dark green ones – resulting from the dominant glass material in mosaics composition since its lucid and compact surface reflects the light in such a way that it is only partly detected by the optical sensors of the instrument. This disturbing phenomenon is often generated in scanning bronze works, which - to overcome this limitation - were, in some cases, treated with powder or spray opacifiers provided their removability. It was considered necessary, in our case, sprinkle the mosaic surface of powder, thus making possible the acquisition of a sufficient amount of points from all the tiles.

In the attempt to compensate for any gap in the data cloud configured in the first scan tests, it was thought to capture the same portion of the surface several times, by inclining the laser emitter at different degrees. For these operations, the manufactured support which allows a shift (50 cm. x 35 cm.) was fixed to a rotatable mechanical base, connected to the computer and operated directly by the software processing of the clouds of points, in order to define precisely the angle of inclination of the surface to be collected with respect to the scanner.

After completing the acquisition of the mosaic, taken to an inclination of 0°, 15°, 30°, 45° and 70°, respectively, it was made the realignment of the five resulting clouds of points. By means of the Polygon Editing Software Tool , supplied as a standard accessory to the scanner by Minolta, one has identified counterparts of the points in each cloud, to be aligned through suitable roto-translations. A subsequent merging feature has produced a single mesh of points, where the lack of information derived from the cloud at 0 ° scan is partially compensated by the superimposition of the other clouds. The final result is shown in Figure 1.

As for the scheduling of a real campaign of three-dimensional relief for mosaics of large dimensions, a few issues that could be solved thanks to the rapid evolution of these technologies have to be carefully planned as previously shown, resulting laser scanning a useful tool for documentation of mosaics. Moreover, reverse engineering methods are interesting for several reasons. The creation of virtual models of small portions, representing compositions where they belong, would permit to carry out

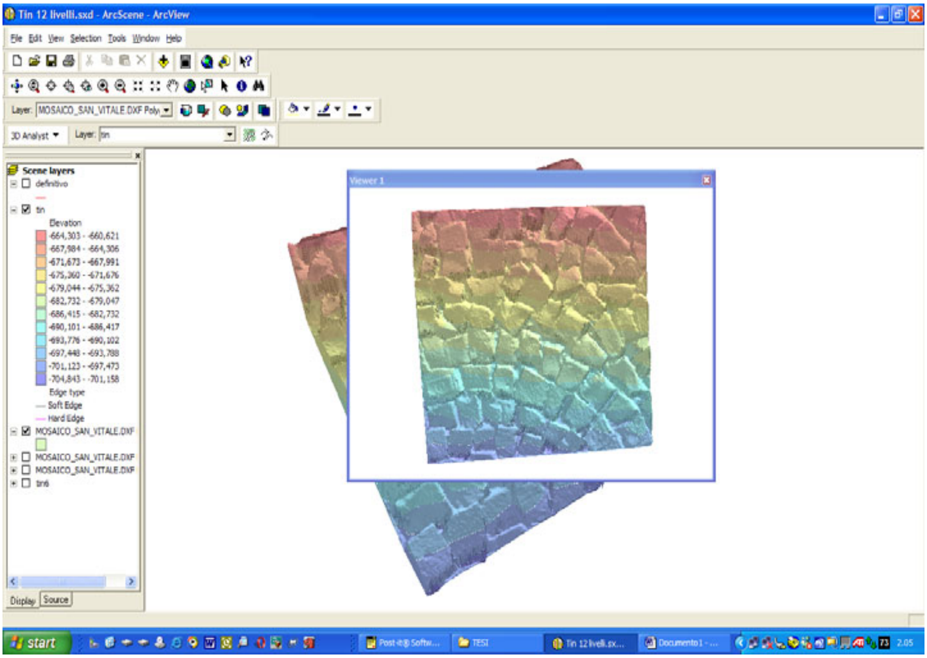
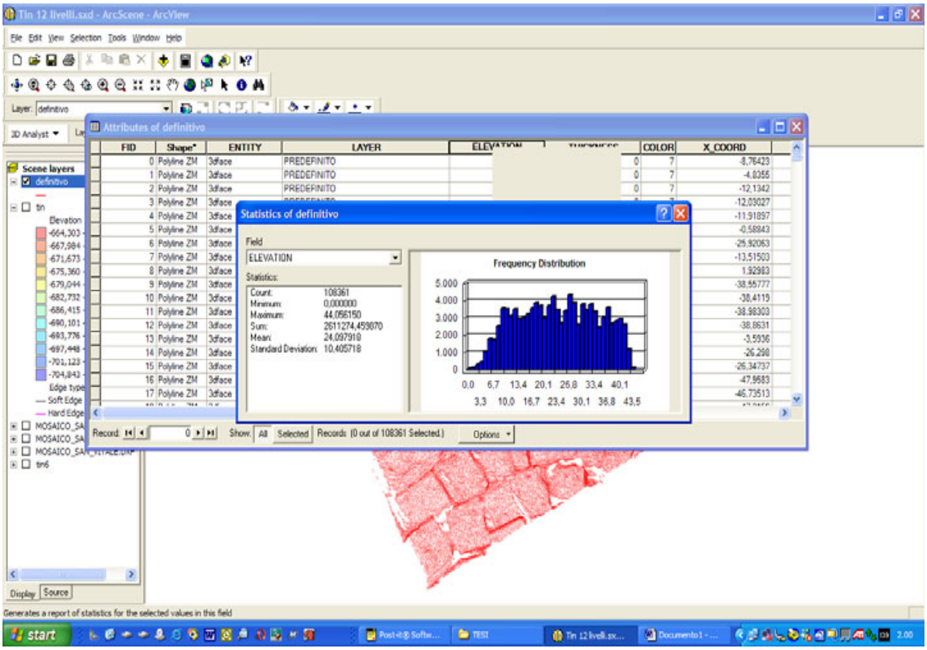


Fig. 1. Three-dimensional rendering of a mosaic sample

comparative investigations of mosaic cycles, leading to quantify similarities and differences in terms of processing and surface rendering of the mosaics themselves. From the virtual model one can perform accurate measurements on the size of the tesserae, their distance, position and orientation, even their projection with respect to the support.

Figure 2 represents the screen view of a geographical information system (GIS) – namely, ArcGis – where the three-dimensional relief of the mosaic has been implemented. Therefore, one can exploit all the features of a GIS software to perform numerical analyses of the whole surface, the position, shape and orientation of single tiles, average values and so on, and create a valuable geo-referenced database documenting in great details the investigated mosaics. As an example, Figure 2 shows a precise measurement of the orientation of a single tessera with respect to the plane of the support.

The application of GIS in a mosaic has been then tested by the integration of data from three-dimensional laser scanning, using software such as Rhinoceros 3.0 and ArcGis 8.01. The first program is a model of land and is used mainly to handle the file format. Stl directly derived from scanning laser data. It has been necessary to convert this format in order to be able to import data into the GIS system, which is not configured to manage the files of this type. The file format .Stl was then converted into a file .Dxf (drawning exchange file), a standard for CAD vector systems. By this way, it was possible to import the model of the mosaic inside ArcGis. The file can then be configured as a set of lines and polygons that describe the geometry of the object and its morphological characteristics.



FID	Shape*	ENTITY	LAYER	ELEVATION	THICKNESS	COLOR	X_COORD
0	Polyline ZM	3dface	PREDEFINITO		0	7	-8,76423
1	Polyline ZM	3dface	PREDEFINITO		0	7	-4,8355
2	Polyline ZM	3dface	PREDEFINITO		0	7	-12,1342
3	Polyline ZM	3dface					-12,03027
4	Polyline ZM	3dface					-11,91897
5	Polyline ZM	3dface					-0,58843
6	Polyline ZM	3dface					-25,92063
7	Polyline ZM	3dface					-13,51503
8	Polyline ZM	3dface					1,92983
9	Polyline ZM	3dface					-38,55777
10	Polyline ZM	3dface					-38,4119
11	Polyline ZM	3dface					-38,98303
12	Polyline ZM	3dface					-38,8631
13	Polyline ZM	3dface					-3,5936
14	Polyline ZM	3dface					-26,298
15	Polyline ZM	3dface					-26,34737
16	Polyline ZM	3dface					-47,9583
17	Polyline ZM	3dface					-46,73513

Record: 0 Show: All Selected Records (0 out of 108361 Selected.) Options

Field: ELEVATION

Statistics:

Count:	108361
Minimum:	0,000000
Maximum:	44,056150
Sum:	2611274,459870
Mean:	24,097918
Standard Deviation:	10,405718

Generates a report of statistics for the selected values in this field

Fig. 2. Screen view of a GIS for mosaics

With these data it has been feasible to build a TIN based on the elevations of geometric primitives. Represented graphically by means of chromatic intervals, the file provides clear and immediate information in the performance of the quantitative measurements of 'tessellato'. The GIS also includes the possibility of statistical analysis on which one can draw in real time characteristics such as maximum and minimum value of the selected fields (height in this case) and the standard deviation, so these precise calculations and immediate morphological study on the surface are easier and quicker than in traditional mappings, thus allowing archaeologist and art historian to have objective data on which to base their thoughts and assumptions on the interpretation of mosaics.

3 Virtual Restoration of Mosaics Cartoons

In the particular case of cartoons representing mosaics, virtual restoration is carried out not in order to prefigure a real restoration, but to enhance and extract information contained in these artworks regarding the grid and the colors of the mosaic with the aim of its reproduction. Therefore, this virtual restoration will not follow the classical criteria of restoration, which would impose the ban to reconstruct the missing parts, although in this case an arbitrary reconstruction is not involved, since it can refer to the original mosaics from where the cartoons were taken from.

Mosaic cartoons are tempera paintings which represent ancient mosaics traced tile by tile. The aim is to create printable replacements of the original cartoons, which may be used in daily teaching activities and which contain as much as possible clear and complete information. In this way the most important cartoons could be taken away and stored in appropriate ways without depriving students from them anyway.

Since specific programs for virtual restoration were not developed, commercial softwares are generally used for vector or bitmaps graphics, in this case, Adobe Photoshop CS. Prior to the real intervention, an adequate CMS system (Color Management System) should be used. Thanks to the use of a colorimeter or a spectrophotometer, CMS can coordinate the gamut, that is the color spaces of different output devices (such as printer and monitor), in order to have the colours of the scanned images correspondent as much as possible to one another and to those of the original artwork.

The process of virtual restoration on the mosaic cartoons consists of the following phases:

♦ digital data acquisition,

♦ mosaicking,

♦ elaboration (balance, cleaning, reconstruction, extraction of the grid),

♦ electronic filing.

Of course, memory of each step and phase of digital processing must be kept in a different level of a Photoshop or GIMP file (with a “psd” extension), able to store several overlying images, saving them without compression and therefore without losing quality.

The digital acquisition is generally performed using a digital camera, the resolution of which today usually ranges from 8 to 14 Megapixels (millions of pixels). To obtain a higher resolution we can also use scanners or digital backs, namely a camera in which the traditional photosensitive film is replaced by a CCD (charge-coupled device) or CMOS (complementary metal-oxide semiconductor) able to capture the image, transforming it into an electrical signal.

If the cartoon to be photographed is very large it can be acquired through several shots, overlapping the edges which then must be reassembled through mosaicking, in order to obtain a resolution of at least 500 or 600 ppi (pixels per inch) on the real size, required to display a screen magnification of all details and to make extremely high quality prints. The mosaicking consists in the perspective rectification of several shots that are then reassembled to form the image of the entire object, perfectly matching the margins.

Once reassembled, not only do the high-resolution photographs allow you to gain a better understanding of the artwork, but they also help to assess the conservation status and identify any previous restoration works.

A first elaboration of the image obtained in this way is to balance brightness and contrast, and to regulate dominant colors, using the "Curves" tool in Photoshop in order to make the picture as similar as possible to the original one, since each acquisition inevitably causes overexposure or underexposure and color toning, due to lighting conditions and the characteristics of the sensors used.

In order to adequately perform these corrections, photos must be carried out by placing a color scale next to the cartoon. By doing so, we can have a reliable reference so as to assess the deviation of color and brightness values of the picture from those of the original one.

For instance in figure 3 we see that the scanned image appears too pale, blurry and slightly turned towards a red color. This is the photograph of the painting by Alessandro Azzaroni representing the pair of doves from the Mausoleum of Galla Placidia. Instead in figure 4 (left) the same image was balanced on the basis of the colors shown by the original mosaic.

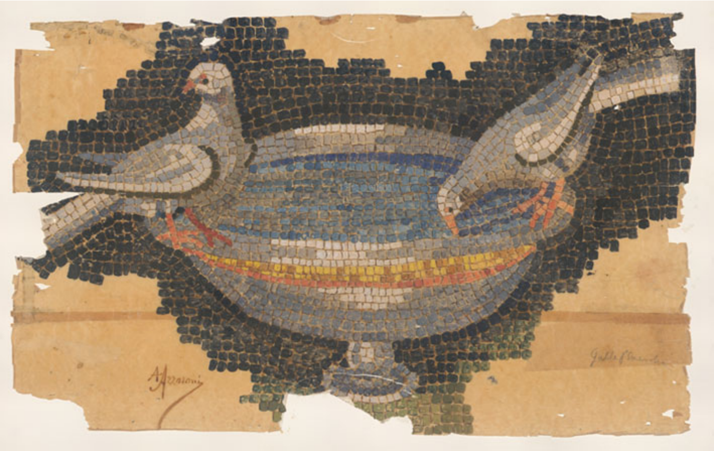


Fig. 3. The painted cartoon by Alessandro Azzaroni representing the pair of doves from the Mausoleum of Galla Placidia

In this way it was possible not only to correct the imperfections due to image acquisition, but also the changes of color of the tesserae, due to alterations of paper color (photochromic degradation) over time; indeed, the new yellow-orange dominant color of paper can be seen through the thin layer of the tempera.

Then, if necessary, you can perform cleanup operations (removal of stains) and intensification of the faded lines. This is a recovery of information which is not always possible to carry out with a traditional restoration intervention. Thanks to the use of IR and UV photography, it is also possible to separate and give back distinctness to any palimpsests.

The reconstruction of missing parts, as we see in figure 4, consists in the use of the image's whole parts in order to recreate the shape of the lost areas, only where it is possible to do so with certainty, therefore for very small gaps and also for much larger ones if there are other copies of the artwork to be restored: in our case we keep the original mosaics on which the cartoons are based.

In addition to the overlapping of different levels corresponding to different phases of work, also the areas of the cartoon where digital reconstruction has been carried out are highlighted by a red edge. The gold tesserae at the base of the cup took on a greenish color in the cartoon, probably due to an alteration of porporina, Pigment consisting of a mixture of metal powders, usually a golden color, easily alterable in contact with air humidity, especially in tempera paintings; in order to give them a color closer to the original we used the "Color replacement" tool.

Conversely, the golden band that decorates the cup was made with yellow, orange and ocher tempera, therefore it did not suffer any alteration.

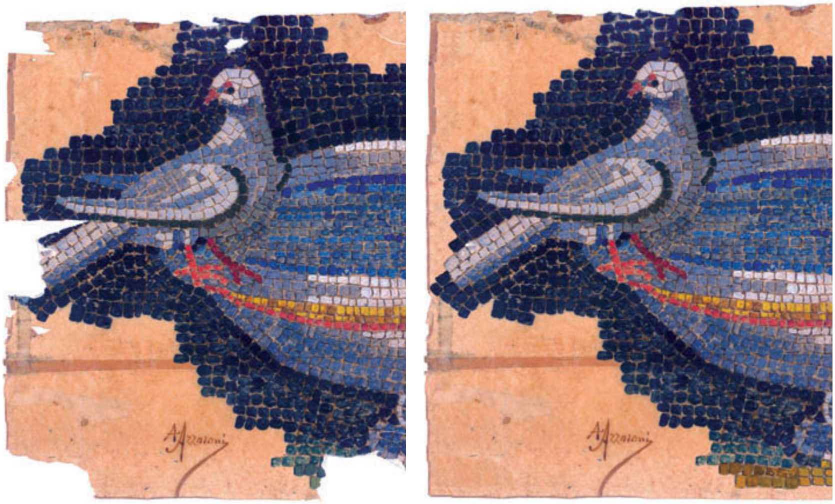


Fig. 4. Drinking doves from Mausoleum of Galla Placidia – cartoon by Alessandro Azzaroni (detail)

The extraction of the grid shown in figure 5, which can be useful to make copies in mosaic, was carried out by using the "Find edges" tool on the image of the restored cartoon, transformed into "Grayscale"; then the contrast was enhanced by using "Curves" to place greater emphasis on the edges of the tesserae, which were then manually cleaned one by one with the "Brush" tool, as the color variations inside were also highlighted, albeit mildly, using the "Find edges" tool.

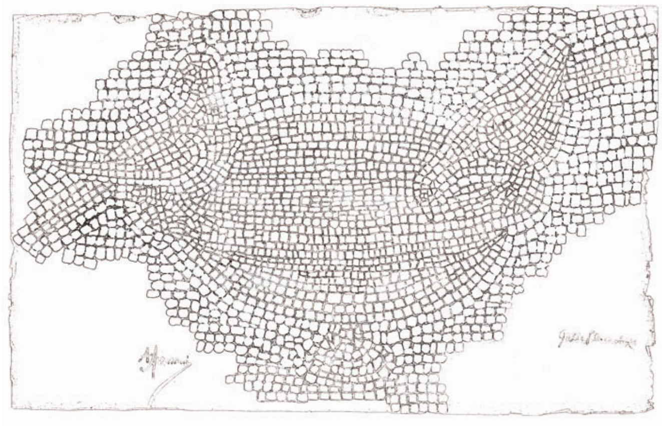


Fig. 5. Extraction of the grid from the mosaic cartoon of figure 4

The last phase of intervention consists in electronic storage of high and low resolution images, in order to keep the acquired images and all subsequent elaborations, foreseeing both their consultation and any new virtual restoration based on different principles. Saved high-resolution images may be used to make high quality prints for periodicals or magazines and for sales. Low resolution ones instead may be used to make on-line catalogues or user-friendly virtual museums. All this should be carried out in order to promote the knowledge of Ravenna's so characteristic heritage formed by mosaics and their cartoons.

During this virtual restoration project, four cartoons representing Ravenna's mosaics belonging to the Severini Institute were photographed.

The images were acquired with a Nikon D70 digital SLR camera, Nikon DX CCD sensor with 6.1 megapixel (23.7x15.6 mm) and a Nikkor AF-S lens (Autofocus-Silent, which uses an SWM technology, Sonic Wave Motor, to make it quieter and faster focusing), and in some cases with the addition of a Hoya 80A filter (light blue), used to correct the prevailing warm tones that often feature pictures taken indoors in artificial light.

Anyhow, any warm and cold tones produced by the filter were subsequently digitally corrected thanks to the color and grey scale used in every shot. The lighting was obtained with the use of flash combined with natural daylight from the window of one of the classrooms in the Severini Institute. The color space used by the camera is sRGB (standard RGB created jointly by HP and Microsoft in 1996) and the size of the photographs is 3008x2000 pixels (25.47x16.93 cm at 300 ppi).

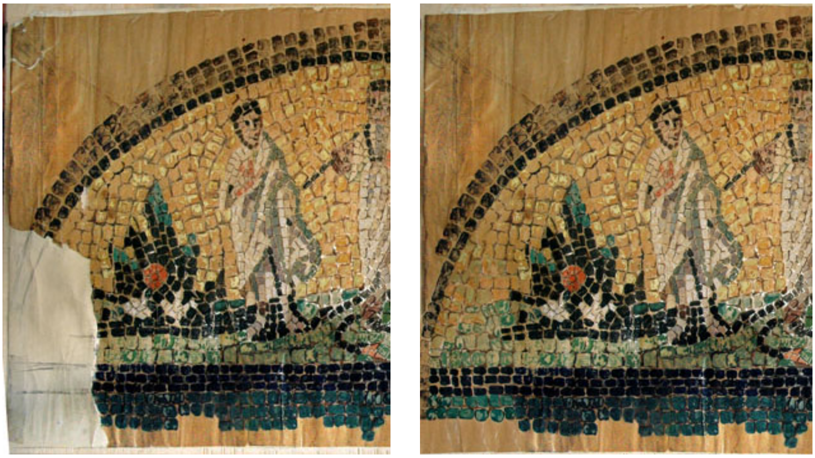


Fig. 6. Lunette with Christ and Apostles from S. Apollinare Nuovo – cartoon by Zelo Molducci (detail).

Given the size (104x70 cm), cartoon N° 54 was photographed in two shots, partially overlapping, that, after color, brightness and contrast balance, were subjected to perspective rectification and mosaicking. The resulting image has a resolution of 90 ppi at actual size. Then the mosaic grid was extracted from the reconstructed image.

So, for photographs of cartoon N° 54 we had to carry out a recomposition process by mosaicking, but not a digital reconstruction intervention of the missing parts, since this picture is preserved intact; on the contrary, cartoons N° 84, 88 and 91, which are smaller, were acquired with only one shot, but since they had different blanks, they were subjected to a mimetic reconstruction of the lost fragments, as shown in figs. 4 and 6.

For image of the cartoon N° 84 we were able to obtain a resolution of 90 ppi at actual size (50x40 cm), while for those of cartoons N° 88 and 91 we have achieved a resolution of 100 ppi at actual size (respectively 68x38.5 cm and 60.8x46.6 cm).

Interventions like these, possibly with the use of tools that improve resolution, carried out on all cartoons with greater historic value, would allow you to preserve the originals from daily use, without subtracting the information contained inside them, but rather improving and spreading them via prints or accessible on-line data banks.

References

[1] Brandi, C.: Teoria del restauro, Edizioni di Storia e Letteratura, 1st edn., Roma (1963); consulted edition: Piccola Biblioteca Einaudi, Torino, pp. 6–8, 26 (1977)

[2] Ferrarini, E., Staltari, E.: Scrittura ed immagini: un'ipotesi di restauro virtuale. Le Médiéviste et l'Ordinateur. Histoire médiévale, informatique et nouvelles technologies (41) (2002)

[3] Bennardi, D., Furferi, R.: Il restauro virtuale. Tra ideologia e metodologia, Edifir, Firenze, p. 13 (2007)

[4] Gonzales, C., Woods, E.: Digital Image Processing. Addison-Wesley, N.Y (1993)

Author Index