LNCS 6854

Pedro Real   Daniel Diaz-Pernil
Helena Molina-Abril   Ainhoa Berciano
Walter Kropatsch (Eds.)

Computer Analysis of Images and Patterns

14th International Conference, CAIP 2011
Seville, Spain, August 2011
Proceedings, Part I

1 Part I

Springer

Lecture Notes in Computer Science 6854

Pedro Real Daniel Diaz-Pernil
Helena Molina-Abril Ainhoa Berciano
Walter Kropatsch (Eds.)

Computer Analysis of Images and Patterns

14th International Conference, CAIP 2011
Seville, Spain, August 29-31, 2011
Proceedings, Part I

Springer

Volume Editors

Ainhoa Bercianol
Universidad del País Vasco
Euskal Herriko Unibertsitatea
Ramón y Cajal, 72, 48014 Bilbao, Spain
E-mail: ainhoa.berciano@ehu.es

Daniel Diaz-Pernil
Helena Molina-Abril
Pedro Real
University of Seville
Avenida Reina Mercedes s/n
41012 Seville, Spain
E-mail: {sbdani, habril, real}@us.es

Walter Kropatsch
Vienna University of Technology
Favoritenstraße 9/186-3
1040 Vienna, Austria
E-mail: krw@prip.tuwien.ac.at

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-23671-6 e-ISBN 978-3-642-23672-3
DOI 10.1007/978-3-642-23672-3
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011935054

CR Subject Classification (1998): I.5, I.4, I.3, I.2.10, I.2.6, H.2.8, I.2.7

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

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

Printed on acid-free paper

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

Preface

This volume contains the papers presented at the 14th International Conference on Computer Analysis of Images and Patterns (CAIP 2011) held in Seville during August 29–31, 2011.

The first CAIP conference was in 1985 in Berlin. Since then CAIP has been organized biennially in different cities around Europe: Wismar, Leipzig, Dresden, Budapest, Prague, Kiel, Ljubljana, Warsaw, Groningen, Versailles, Vienna and Münster.

Following the spirit of the previous meetings, the 14th CAIP was conceived as a period of active interaction among the participants, with emphasis on exchanging ideas and on cooperation.

This year, 286 full scientific papers from 52 countries were submitted, of which 138 were accepted for presentation based on the positive scientific reviews. All the papers have been revised by, at least, two reviewers and, most of them by three.

The accepted papers were presented during the conference either as oral presentations or as posters in the single-track scientific program. Oral presentations allowed the authors to reach a large number of participants, while posters allowed for a more intense scientific interaction. We tried to continue the tradition of CAIP in providing a forum for scientific exchange at a high-quality level.

Two internationally recognized speakers accepted our invitation to present a stimulating research topic this year: Peter Sturm, INRIA Grenoble (France) and Facundo Memoli, Stanford University (USA).

Indeed, these proceedings are divided into two volumes, 6854 and 6855, where the index has been structured following the topics and program of the conference.

We are grateful for the great work realized by the Program Committee and additional reviewers. We especially thank the PRIP and CATAM members, who made a big effort to help.

We appreciate our sponsors for their direct and indirect financial support and Springer for giving us the opportunity to continue publishing CAIP proceedings in the LNCS series.

Finally, many thanks go to our local support team and, mainly, to María José Jiménez Rodríguez for her huge and careful work of supervision of almost all the tasks of the Organizing Committee.

August 2011

Ainhoa Berciano
Daniel Diaz-Pernil
Walter Kropatsch
Helena Molina-Abril
Pedro Real

CAIP 2011 Organization

Conference Chairs

Pedro Real	University of Seville, Spain
Walter Kropatsch	Vienna University of Technology, Austria

Steering Committee

André Gagalowicz (France)
Xiaoyi Jiang (Germany)
Reinhard Klette (New Zealand)
Walter Kropatsch (Austria)
Nicolai Petkov (The Netherlands)
Gerald Sommer (Germany)

Program Committee

Shigeo Abe
Ceyhun Burak Akgul
Mayer Aladjem
Sylvie Alayrangues
Madjid Allili
A. Antonacopoulos
Heider Araujo
Jonas August
Antonio Bandera
Elisa H. Barney Smith
Brian A. Barsky
Algirdas Bastys
E. Bayro Corrochano
Ardhendu Behera
Abdel Belaid
Olga Bellon
Ainhoa Berciano
Wolfgang Birkfellner
Dorothea Blostein
Gunilla Borgefors
Christian Breiteneder
Thomas Breuel
Luc Brun
Lorenzo Bruzzone
Martin Burger
Gustavo Carneiro
Kwok Ping Chan
Yung-Kuan Chan
Rama Chellappa
Sei-Wang Chen
Da-Chuan Cheng
Dmitry Chetverik
Jose Cortes Parejo
Bertrand Couasnon
Marco Cristani
Guillaume Damiand
Justin Dauwels
Mohammad Dawood
Gerard de Haan
Alberto Del Bimbo
Andreas Dengel
Joachim Denzler
Cecilia Di Ruberto
Daniel Diaz-Pernil
Philippe Dosch
Hazim Kemal Ekenel
Neamat El Gayar
Hakan Erdogan
Francisco Escolano
M. Taner Eskil
Chiung-Yao Fang
Miguel Ferrer
Massimo Ferri
Gernot Fink
Robert Fisher
Ana Fred
Patrizio Frosini
Laurent Fuchs
Xinbo Gao
Anarta Ghosh
Georgy Gimel'farb
Dmitry Goldgof
Rocio Gonzalez-Diaz
Cosmin Grigorescu
M.A. Gutierrez-Naranjo
Michal Haindl
Edwin Hancock
Changzheng He
Vaclav Hlavac
Zha Hongbin
Joachim Hornegger
Yo-Ping Huang
Yung-Fa Huang
Atsushi Imiya
Shuiwang Ji
Xiaoyi Jiang
Maria Jose Jimenez
Martin Kampel
Nahum Kiryati
Reinhard Klette
Andreas Koschan

Walter Kropatsch
James Kwok
Longin Jan Latecki
Xuelong Li
Pascal Lienhardt
Guo-Shiang Lin
Josep Llados
Jean-Luc Mari
Eckart Michaelse
Ioana Necula
Radu Nicolescu
Mads Nielsen
Darian Onchis-Moaca
Samuel Peltier
Petra Perner
Nicolai Petkov
Ioannis Pitas
Mario J. Perez Jimnez
Petia Radeva
Pedro Real
Jos Roerdink
Bodo Rosenhahn
Jose Ruiz-Shulcloper
Robert Sablatnig
Robert Sabourin
Hideo Saito
Albert Salah
Gabriella Sanniti Di Baja
Sudeep Sarkar
Oliver Schreer
Francesc Serratosa
Luciano Silva
Gerald Sommer
Mingli Song
K.G. Subramanian
Akihiro Sugimoto
Dacheng Tao
Klaus Toennies
Karl Tombre
Javier Toro
Andrea Torsello
Chwei-Shyong Tsai
Ernest Valveny
Mario Vento
Jose Antonio Vilches
Steffen Wachenfeld
Shengrui Wang
Michel Westenberg
Paul Whelan

Additional Reviewers

Nicole Artner
Facundo Bromberg
Christoph Brune
Javier Carnero
Andrea Cerri
Chao Chen
Wen-Chang Cheng
Michel Devy
Denis Enachescu
Yll Haxhimusa
Chih-Yu Hsu
Adrian Ion
Jiun-Jian Liaw
Helena Molina-Abril
Gennaro Percannella
Federico Schluter
Cheng-Ying Yang
Chih-Chia Yao

Local Organizing Committee

Ainhoa Berciano
Javier Carnero
Daniel Diaz-Pernil
Maria Jose Jimenez
Ioana Necula
Belen Medrano
Helena Molina-Abril
Ana Pacheco
Regina Poyatos
Angel Tenorio
Lidia de la Torre

Sponsoring Institutions

Vicerrectorado de Investigación, Universidad de Sevilla
Instituto de Matemáticas de la Universidad de Sevilla, A. de Castro Brzezicki
Fundación para la Investigación y el Desarrollo de las Tecnologías de la Información en Andalucía
Ministerio de Ciencia e Innovación (Spain)
Consejería de Economía, Ciencia e Innovación de la Junta de Andalucía
International Association for Pattern Recognition (IAPR)
Escuela Técnica superier de Ingeniería Informática, Universidad de Seville, Spain
Department of Applied Mathematics I, University of Seville, Spain

Table of Contents – Part I

Invited Lecture

Motion Analysis

Image and Shape Models

Segmentation and Grouping

Shape Recovery

Graph-Based Methods and Representations

Curves, Surfaces and Objects beyond 2 Dimensions

Geo-topological Analysis of Images

Kernel Methods

Image and Video Indexing and Database Retrieval

Object Detection and Recognition

Medical Imaging

Structural Pattern Recognition

Table of Contents – Part II

Invited Lecture

Biometrics

Human and Face Detection and Recognition

Document Analysis

Applications

3D Vision

Image Restoration

Restoration

Natural Computation for Digital Imagery

Image and Video Processing

Calibration

Color and Texture

Tracking and Stereo Vision

A Historical Survey of Geometric Computer Vision

Peter Sturm

INRIA Grenoble Rhône-Alpes and Laboratoire Jean Kuntzmann, Grenoble, France
Peter.Sturm@inria.fr

Abstract. This short paper accompanies an invited lecture on a historical survey of geometric computer vision problems. It presents some early works on image-based 3D modeling, multi-view geometry, and structure-from-motion, from the last three centuries. Some of these are relatively well known to photogrammetrists and computer vision researchers whereas others seem to have been largely forgotten or overlooked. This paper gives a very brief summary of an ongoing historical study.

Keywords: Geometry, photogrammetry, structure-from-motion, 3D modeling, history.

1 3D Modeling from Perspectives

Relatively quickly after the invention of photography in the 1830's[1], the idea of using photographs for map creation and 3D modelling (terrains, buildings) emerged. Laussedat seems to have been the first to exploit photographs for topographic modelling [32]; he is thus commonly considered as the father of photogrammetry. In the 1850's and 1860's, he developed and used an approach for topographic map generation, first with perspective drawings generated using the so-called *camera lucida* [30], later with actual photographs. The approach requires perspective images taken in particular conditions (horizontal optical axes, known distance between the viewpoints, known internal parameters) and carries out the mapping using elementary operations for the determination of the relative horizontal orientation of the perspectives, of planar point triangulation, and height measurement. The cameras were usually coupled with theodolites, providing accurate angle measurements.

It is noteworthy that essentially the same principles were used already before the existence of photography to generate topographic maps. Beautemps-Beaupré developed an approach to do so, from hand-drawn "perspectives" to which angles measured by theodolites, between sight lines to pairs of points, were added [2]. Here, the hand-drawn images seem mainly to have been used for documentation purposes as opposed to for actual measurements. This approach was used in marine expeditions as early as in the 1790's to acquire topographic maps of

[1] The invention of photography was the result of long-term efforts of many researchers in different countries, although the "official" birthdate is often given as 1837 or 1839.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 1–8, 2011.

foreign coastlines, from observations made on board of ships and where possible, combining these with land-based measurements.

An even earlier example of 3D modeling aided by drawings may be due to Kappeler[2], who produced in 1726 a topographic map of a mountain range in Switzerland, Mount Pilatus [24,13]. Although Kappeler stated that he used perspective drawings, no details on the approach are given, which is why in photogrammetric litterature this work is usually not considered as the first undisputable example of 3D modeling using inverse perspective methods.

2 “Hardware”

The maturation of photogrammetry was made possible by theoretical developments, as well as, if not more, by practical inventions. The latter concerned of course photographic equipment as such, but also devices that ease the use of cameras for measurement purposes, such as combinations of cameras and theodolites. While such “hardware developments” are not the central issue of this paper, we still like to mention that concepts such as panoramic image acquisition and multi-camera systems, were developed relatively early. Indeed, it was soon recognized that panoramic images may ease photogrammetric work; the first panoramic camera may be one developed by Puchberger in 1843, i.e. just a few years after the invention of modern photography [35]. Multi-camera systems were developed at least as early as in 1884, initially mainly if not exclusively for aerial imaging. The earliest work known to me (no effort was made for an exhaustive bibliography research) is that of Triboulet, who, as reported in [45] experimented from 1884 on with a multi-camera system consisting of 7 cameras attached to a balloon: one camera looked downwards and 6 cameras were equally distributed around the balloon’s circumference (the system thus resembles the popular Ladybug sensor). In addition to hardware for the acquisition of images and complementary measurements, photogrammetry progressed signficantly through the development of equipment and procedures to exploit the acquired images, see e.g. an early survey in [31,33].

3 Epipolar and Multi-view Geometry

Epipolar geometry seems to have been first uncovered by Hauck in 1883 [17]. In the same paper as well as follow-up papers [18,19,20,21], trilinear relationships of points and lines seen in three images, were also described. In his work, Hauck did not deeply analyze these trilinear relationships theoretically, like was done via trifocal tensors in the 1990’s; he rather concentrated on the application of these relationships to generate a third image from two given ones (often called “trifocal transfer” in computer vision).

Previously, in 1863, Hesse proposed an algebraic solution to an exercise proposed by Chasles in [3]: given seven pairs of matching 2D points, the goal is to

[2] Often spelled Cappeler.

determine two pencils of 2D lines in homographical relation such that matching lines are incident with matching points. Further, Chasles asked to prove that there exist only three solutions to this problem. Hesse proposed a solution to this problem that is effectively equivalent to the 7-point method for computing the fundamental matrix, or epipolar geometry, between two perspective images of a rigid 3D object [22], although the link to epipolar geometry only became clear later.

Before Hesse, de Jonquières gave a geometrical solution and proof for Chasles' problem [7], which were later slightly clarified by Cremona [4]. Sturm studied the same problem as well as generalizations to other numbers of points, from a theoretical viewpoint [41].

The special case of six point matches for which it is known that four arise from coplanar points in the scene, was solved by Finsterwalder in 1899 [10].

4 Projective Reconstruction

Hauck, in the above works, already touches upon the issue of projective reconstruction [17,18,19,20,21]. In 1899, Finsterwalder gives a clear exposition of the fact that from a set of uncalibrated images, a projective 3D reconstruction is possible and provides an algorithm for the case of two images [10]. The concept of projective reconstruction was re-discovered in computer vision in the early 1990's [9,16].

5 Self-calibration

In the same article where he explained the feasibility of projective reconstruction, Finsterwalder also showed that self-calibration from images of an unknown rigid object is possible [10]. He proposed a geometric construction based on the absolute conic and the circular points of image planes, didn't find an analytical solution though. In computer vision, self-calibration was first formulated for the case of images acquired by a camera with fixed intrinsic parameters [34] and then extended towards images acquired with different intrinsics [23,38]. Finsterwalder directly considered the problem of images with possibly different intrincis; he showed that with four images taken with possibly varying focal length and principal point, a finite number of solutions exists.

His construction goes as follows (formulated in computer vision jargon, the original explanations being somewhat different). Given a projective reconstruction of the object and the cameras. Under the assumption of square pixels, the image of the absolute conic is a circle, hence it contains the two circular points of the image plane. Let us now back-project the circular points to 3D lines in the projective reconstruction; these must intersect the absolute conic. For four images, we thus get eight 3D lines that intersect the absolute conic. This is in general sufficient to determine (a finite number of solutions for) the plane at infinity (3 degrees of freedom) and the absolute conic lying on it (5 degrees of freedom). Hence, the self-calibration problem can in general be solved.

Self-calibration from rotations. The possibility of determining the focal length from images acquired by rotating a camera, was mentioned by Meydenbauer in 1892 [36]. In [46], von Gruber described topographic and photogrammetric acquisitions made during an expedition in the Pamir mountain range in 1913 and subsequent measurements and the generation of a map. Panoramic image acquisition by rotating the camera about its optical center, was an integral part of the procedures used. It is mentioned that this was also used to determine the focal length of cameras.

The above approaches deal with the case of known principal point and "only" determine the focal length [42]. This was generalized in 1939 by Sutor to the determination of both, the focal length and the principal point, as well as radial distortion, from a set of images spanning a full circle [42]. His method, like the others above, assumes that the camera rotates about either the horizontal or vertical axis of its coordinate system (errors resulting from deviations from this setup are investigated by Sutor and shown to be negligible in practice). Thus, only the horizontal or vertical coordinate of the principal point is considered respectively computed. Sutor's method is iterative, starting from initial values. Besides providing details of the method, Sutor also gave a theoretical error analysis.

Wester-Ebbinghaus extended this approach towards using images acquired in arbitrary orientations around the fixed optical center and without requiring a closed image sequence [48]. Like Sutor, he did not propose "closed-form" solutions and solved the problem in a bundle adjustment manner. He also proposed a bundle adjustment formulation for the case where a camera rotates about a fixed point different from the optical center.

General closed-form solutions were developed by Hartley et al. [15,6].

(Self-)calibration from 3 orthogonal vanishing points. This is a simple (in general inaccurate) calibration idea which only requires one image of a rectangular parallelepiped and the extraction of the 3 vanishing points associated with its edges. Such an image allows to compute the camera's focal length and principal point (if the object is in general position). This idea was discovered independently by many researchers over time, starting with the famous mathematician Taylor in 1715 [43]. Later references include [10,11,8,1].

6 Orientation Procedures – Structure-from-Motion

The main building blocks for structure-from-motion algorithms are what is called orientation procedures in photogrammetry: pose estimation (space resection), motion estimation (relative orientation), and 3D point triangulation (intersection). These are classical problems, see e.g. the excellent overview [49]. A few notes on pose and motion estimation, follow.

Pose estimation. Taylor and Lambert were probably among the first to have posed "inverse perspective problems" in a general manner [43,44,29], both in the 18th century. Lambert (who also studied shading and developed the so-called

Lambertian reflectance model), proposed and solved a series of such problems, including pose estimation, estimation of the focal length, and of the orientation of an image. This was mostly restricted to special cases, for instance the assumption of an image plane that is orthogonal to a ground plane and the existence of rectangles or squares on the ground plane [29]. Interestingly, Lambert's studies were partly motivated by aesthetic considerations: his premise was that in order to best contemplate a (perspective) painting, the observer should put himself in a position relative to the painting that corresponds to the painter's eyepoint relative to the depicted scene. His initial goal thus seems to have been the determination of this "pose" from information contained in the painting (e.g. vanishing points) and additional knowledge.

Maybe the first analytical solution to the so-called 3-point pose problem, was found by Lagrange: determine the position and orientation of a "camera" (Lagrange obviously didn't speak of cameras) relative to three 3D points, given knowledge of the relative positions of these points and the angles spanned by pairs of points and the optical center. Lagrange discussed this problem at least as early as from 1773 on [27,28]. He already showed that it can be reduced to finding the roots of a quartic polynomial and also sketches an iterative numerical procedure. He very likely had the complete solution although the above publications only give a general sketch and do not contain all details.

A complete analytical solution was eventually given by Grunert in 1841 [12]. Many other solutions have been proposed in the literature since, see e.g. the survey [14].

Motion estimation. Kruppa showed in 1913 that from five point matches between two calibrated images of a rigid object, the relative pose between the images can be computed up to a finite number of solutios [26]. Later works on the 5-point problem are [37] and references therein.

The special case of a planar object was already solved by Schröter in 1880 (paragraph 45 of [39]). He showed how from two calibrated images of four coplanar points, these points as well as the camera positions and orientations, can be computed up to two solutions. An equivalent result was given later by Kruppa (section A of [25]).

7 Special Cases of 3D Modeling

Shape from silhouettes. Amazingly, this was one of the first 3D modeling approaches to be developed: around 1857, François Willème developed an approach, baptized *photo-sculpture* (see e.g. [47]), that is nothing else than a "mechanical" version of shape from silhouettes or, the visual hull. Willème acquired images in a circle around an object (usually, a person). These were then projected to a screen, one after the other; behind the screen, a block of clay or other material easy to sculpt, was positioned on a turntable, which was rotated in order to reproduce the current image's orientation while it was acquired. Then, using a so-called pantograph (an articulated instrument), the sculptor followed the silhouettes of the object on the projector screen, while steering a bar that carved away the parts of the clay block that lie outside the silhouette. By repeating this procedure for all images, the outcome is nothing else than the object's visual hull in clay! This was finally worked on by an actual sculptor, to round edges and add details, in order to produce a visually pleasing statue. Interestingly, this concept seems to have been very popular in high society circles for several years, and Willème created what one might nowadays call a start-up company, commercializing this concept (he was even imitated in other countries).

Single-view 3D modeling. The idea of performing 3D modeling from a single image of an object, was proposed by several researchers in the late 19th century [17,10,36,32]. Like their modern counterparts, see e.g. [5,40], the proposed approaches relied on the exploitation of geometric constraints provided by the user, such as parallelism of lines, right angles, etc.

8 Conclusion

As mentioned in the abstract, this paper is the result of a historical study that is in progress. In the mid-term future, it will be complemented by a more complete treatment, containing more technical details, (many) more references, and covering other aspects such as 3D modeling from shadows, structure-from-motion for refractive objects (also called "multimedia-photogrammetry"), structured lighting, 3D modeling of surfaces of revolution, etc.

Acknowledgement. I am grateful to the Alexander-von-Humboldt Foundation for a Research Fellowship supporting a sabbatical at TU Munich, during which the literature search underpinning this survey was mainly carried out.

References

1. Caprile, B., Torre, V.: Using Vanishing Points for Camera Calibration. IJCV 4, 127–140 (1990)
2. Chapuis, O.: À la mer comme au ciel – Beautemps-Beaupré et la naissance de l'hydrographie moderne (1700-1850). Presse de l'Université de Paris-Sorbonne (1999)
3. Chasles, M.: Question 296. Nouvelles annales de Mathématiques 1(14), 50 (1855)
4. Cremona, M.: Sur un problème d'homographie (question 296). Nouvelles annales de Mathématiques 1(20), 452–456 (1861)
5. Criminisi, A.: Accurate Visual Metrology from Single and Multiple Uncalibrated Images. PhD thesis, University of Oxford (1999)
6. de Agapito, L., Hayman, E., Hartley, R.I.: Linear self-calibration of a rotating and zooming camera. In: CVPR, pp. 1015–1021 (1999)
7. de Jonquières, M.: Solution géométrique de la question 296. Nouvelles annales de Mathématiques 1(17), 399–403 (1858)
8. Echigo, T.: A camera calibration technique using three sets of parallel lines. Machine Vision and Applications 3(3), 159–167 (1990)
9. Faugeras, O.: What can be seen in three dimensions with an uncalibrated stereo rig? In: Sandini, G. (ed.) ECCV 1992. LNCS, vol. 588, pp. 563–578. Springer, Heidelberg (1992)

10. Finsterwalder, S.: Die geometrischen Grundlagen der Photogrammetrie. Jahresbericht Deutscher Mathematik 6, 1–44 (1899)
11. Gracie, G.: Analytical photogrammetry applied to single terrestrial photograph mensuration. In: XIth International Congress of Photogrammetry (1968)
12. Grunert, J.A.: Das pothenot'sche Problem in erweiterter Gestalt; nebst Bemerkungen über seine Anwendung in der Geodäsie. Archiv der Mathematik und Physik, 238–248 (1841)
13. Günther, L.W.: Die erste praktische Anwendung des Meßbildverfahrens durch den Schweizer M.A. Cappeler im Jahre 1725. International Archives of Photogrammetry 3(4), 289–290 (1913)
14. Haralick, R.M., Lee, C., Ottenberg, K., Nölle, M.: Analysis and solutions of the three point perspective pose estimation problem. In: CVPR, pp. 592–598 (1991)
15. Hartley, R.I.: An algorithm for self calibration from several views. In: CVPR, pp. 908–912 (1994)
16. Hartley, R.I., Gupta, R., Chang, T.: Stereo from uncalibrated cameras. In: CVPR, pp. 761–764 (1992)
17. Hauck, G.: Neue Constructionen der Perspective und Photogrammetrie (Theorie der trilinearen Verwandtschaft ebener Systeme) – 1st article. Journal für die reine und angewandte Mathematik 95(1), 1–35 (1883)
18. Hauck, G.: Theorie der trilinearen Verwandtschaft ebener Systeme. die orientirte Lage – 2nd article. Journal für die reine und angewandte Mathematik 97(4), 261–276 (1884)
19. Hauck, G.: Theorie der trilinearen Verwandtschaft ebener Systeme. die dreibündig-eindeutige Verwandtschaft zwischen drei ebenen Punktsystemen und ihre Beziehungen zur quadratischen und zur projectiv-trilinearen Verwandtschaft – 3rd article. Journal für die reine und angewandte Mathematik 98(4), 304–332 (1885)
20. Hauck, G.: Theorie der trilinearen Verwandtschaft ebener Systeme. die trilineare Beziehung zwischen drei einstufigen Grundgebilden – 4th article. Journal für die reine und angewandte Mathematik 108(1), 25–49 (1891)
21. Hauck, G.: Theorie der trilinearen Verwandtschaft ebener Systeme. zusammenfassung und wichtige Specialfälle – 5th article. Journal für die reine und angewandte Mathematik 111(3), 207–233 (1893)
22. Hesse, O.: Die cubische Gleichung, von welcher die Lösung des Problems der Homographie von M. Chasles abhängt. Journal für die reine und angewandte Mathematik 62, 188–192 (1863)
23. Heyden, A., Åström, K.: Euclidean reconstruction from image sequences with varying and unknown focal length and principal point. In: CVPR, pp. 438–443 (1997)
24. Kappeler, M.A.: Pilati Montis Historia. In: Imhof, J.R., Filii (eds.) Pago Lucernensi Helvetiae Siti, Figuris Aeneis Illustrata, Basel, Switzerland (1767)
25. Kruppa, E.: Über einige Orientierungsprobleme der Photogrammetrie. Sitzungsberichte der mathematisch-naturwissenschaftlichen Klasse der kaiserlichen Akademie der Wissenschaften, Abteilung II a 121(1), 3–16 (1912)
26. Kruppa, E.: Zur Ermittlung eines Objektes aus zwei Perspektiven mit innerer Orientierung. Sitzungsberichte der mathematisch-naturwissenschaftlichen Klasse der kaiserlichen Akademie der Wissenschaften, Abteilung II a 122(10), 1939–1948 (1913)
27. Lagrange, J.-L.: Solutions analytiques de quelques problèmes sur les pyramides triangulaires (1773); reprinted in 1869 in the 3rd volume of the Œuvres de Lagrange edited by J.-A. Serret and published by Gauthier-Villars

28. Lagrange, J.-L.: Leçons élémentaires sur les mathématiques données à l'École Normale en (1795); 1795, reprinted in 1877 in the 7th volume of the Œuvres de Lagrange, edited by J.-A. Serret and published by Gauthier-Villars
29. Lambert, J.H.: Die freye Perspektive, oder Anweisung, jeden perspektivischen Aufriß von freyen Stücken und ohne Grundriß zu verfertigen. Heidegger und Compagnie, Zurich (1759)
30. Laussedat, A.: Mémoire sur l'emploi de la chambre claire dans les reconnaissances topographiques, Mallet-Bachelier, Paris. Mémorial du Génie, vol. 16 (1854)
31. Laussedat, A.: Recherches sur les instruments, les méthodes et le dessin topographiques, vol. 1. Gauthier-Villars, Paris (1898)
32. Laussedat, A.: La métrophotographie. Gauthier-Villars, Paris (1899)
33. Laussedat, A.: Recherches sur les instruments, les méthodes et le dessin topographiques, vol. 2. Gauthier-Villars, Paris (1901)
34. Maybank, S.J., Faugeras, O.D.: A theory of self calibration of a moving camera. IJCV 8(2), 123–151 (1992)
35. McBride, B.: A timeline of panoramic cameras, http://www.panoramicphoto.com/timeline.htm?
36. Meydenbauer, A.: Das photographische Aufnehmen zu wissenschaftlichen Zwecken, insbesondere das Messbild-Verfahren – Erster Band: Die photographischen Grundlagen und das Messbild-Verfahren mit kleinen Instrumenten. Unte's Verlags-Anstalt, Berlin (1892)
37. Nistér, D.: An efficient solution to the five-point relative pose problem. IEEE-TPAMI 26(6), 756–770 (2004)
38. Pollefeys, M., Koch, R., Van Gool, L.: Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters. In: ICCV, pp. 90–95 (1998)
39. Schröter, H.: Theorie der Oberflächen zweiter Ordnung und der Raumkurven dritter Ordnung als Erzeugnisse projektivisher Gebilde. In: Nach Jacob Steiner's Prinzipien auf synthetischem Wege abgeleitet. B.G. Teubner, Leipzig (1880)
40. Sturm, P., Maybank, S.J.: A method for interactive 3D reconstruction of piecewise planar objects from single images. BMVC, 265–274 (1999)
41. Sturm, R.: Das Problem der Projektivität und seine Anwendung auf die Flächen zweiten Grades. Mathematische Annalen 1, 533–574 (1869)
42. Sutor, J.: Bestimmung der inneren Orientierung und Verbildung aus Rundbildern. PhD thesis, Technische Hochschule Berlin (1939)
43. Taylor, B.: Linear perspective: or, a new method of representing justly all manner of objects as they appear to the eye in all situations (1715)
44. Taylor, B.: New principles of linear perspective: or the art of designing on a plane the representations of all sorts of objects, in a more general and simple method than has been done before (1719)
45. Tissandier, G.: La photographie en ballon. Gauthier-Villars (1886)
46. von Gruber, O.: Topographische Ergebnisse der Pamir-Expedition des D. u. ö (Deutschen und österreichischen) Alpenvereines 1913. International Archives of Photogrammetry 6, 156–181 (1923)
47. Wells, D.A.: Photo-sculpture. Annual of Scientific Discovery: Or, Year-Book of Facts in Science, 183–185 (1865)
48. Wester-Ebbinghaus, W.: Einzelstandpunkt-Selbstkalibrierung – ein Beitrag zur Feldkalibrierung von Aufnahmekammern, Habilitation thesis (1982)
49. Wrobel, B.: Minimum solutions for orientation. In: Gruen, A., Huang, T.S. (eds.) Calibration and Orientation of Cameras in Computer Vision. Springer, Heidelberg (2001)

Detection Human Motion with Heel Strikes for Surveillance Analysis

Sung-Uk Jung and Mark S. Nixon

ISIS, School of Electronics and Computer Science
University of Southampton, SO17 1BJ, UK
{suj08r,msn}@ecs.soton.ac.uk

Abstract. Heel strike detection is an important cue for human gait recognition and detection in visual surveillance since the heel strike position can be used to derive the gait periodicity, stride and step length. We propose a novel method for heel strike detection using a gait trajectory model, which is robust to occlusion, camera view and to low resolution which can generalize to a variety of surveillance imagery. When a person walks, the movement of the head is conspicuous and sinusoidal. The highest point of the trajectory of the head occurs when the feet cross. Our gait trajectory model is constructed from trajectory data using non-linear optimization. Then, the key frames in which the heel strike takes place are extracted. A Region Of Interest (ROI) is extracted using the silhouette image of the key frame as a filter. Finally, gradient descent is applied to detect maxima which are considered to be the time of the heel strikes. The experimental results show a detection rate of 95% on two databases. The contribution of this research is the first use of the gait trajectory in the heel strike position estimation process and we contend that the approach is a new approach for basic analysis in surveillance imagery.

Keywords: Heel strike detection, gait trajectory model, gradient descent, gait.

1 Introduction

Heel strike detection is a basic and important process in non-invasive analysis of people at a distance, especially, for human motion analysis, and recognition by gait, in a visual surveillance environment. Since the gait periodicity, stride and step length can be calculated directly from the position of the heel strikes; this information can be used to represent the individual characteristics of a human, the walking direction, and the basic 3D position information (given a calibrated camera).

There are two central observations concerning heel strike detection. During the strike phase, the foot of the striking leg stays at the same position for half a gait cycle (when the foot is in contact with the floor), whilst the rest of the human body moves forward [1]. Another is that when the left foot and right feet cross, the head is at its highest position in a gait cycle. We develop our new heel strike detection method based on these observations.

Generally, heel strike detection is a preliminary step in gait recognition or model-based human body analysis and visualization. There are two major previous

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 9–16, 2011.

approaches to heel strike detection. The first is a model-free approach which uses low level data such as silhouettes and edges to detect gait motion. Bobick and Johnson [2] recovered static body and stride parameters of subjects using the action of walking to extract relative body parameters. BenAbdelkader et al. [3] identified people from low resolution video by estimating the height and stride parameters of their gait. Jean et al. [4] proposed an automatic method of detecting body parts using a human silhouette image. A five point human model was detected and tracked. They solved for self-occlusion of the feet by using optical flow and motion correspondence. Bouchrika and Nixon [1] built the accumulator map of all of corner points using Harris corner detector during an image sequence. Then, the heel strike position was estimated using the density of proximity of the corner points.

An alternative approach is a model-based approach which uses prior information such as 3D shape, position and trajectory of body motion. The heel strike position is sub-result of these researches. Vignola et al. [5] fitted a skeleton model to a silhouette image of person. Each limb (two arms, two legs) was fitted independently to speed-up the fitting process. Zhou et al. [6] extracted full-body motion of walking people from video sequences. They proposed a Bayesian framework to introduce prior knowledge into system for extracting human gait. Sigal and Black [7] estimated human pose using occlusion-sensitive local image likelihoods method. Zhang et al. [8] presented a 3-level hierarchical model for localizing human bodies in still images from arbitrary viewpoints. They handled self-occlusion and large viewpoint changes using Sequential Monte Carlo (SMC) optimization. Sundaresan et al. [9] proposed a graphical model of the human body to segment 3D voxel data of human into different articulated chains.

However, many approaches consider the fronto-parallel view of a walking subject where the subject walks in a direction normal to the camera's plane of view. Also, in the model-based approaches there is much computational load in initialization and tracking. Moreover, in the visual surveillance environment the image quality could be low and the image sequences derived from a single camera only are available. Therefore, an alternative heel strike detection method is needed which is robust to low resolution, foot self-occlusion, camera view point and suitable for single camera.

In this paper, to overcome the above constraints, we propose a novel method of heel strike detection using the gait trajectory model. As mentioned before, the frame in which the heel strike takes place can be extracted through the gait motion even when the foot is hidden by another leg and the image quality is low. In this research, the gait trajectory model [10] is deployed. This model is applied to detect the key frame in which the heel strikes happen. The silhouette image at the key frame is used to remove background data.

The remainder of this paper is organized as follows: Section 2 explains the key frame calculation. Section 3 describes the heel strike candidate detection and verification method using the key frame and gradient descent. Section 4 shows the experimental results based on visual surveillance databases. In Section 5, we conclude our work.

2 Key Frame Calculation

To find the moment of a heel strike the basic characteristics of gait is used. When a person walks the movement of head is conspicuous and sinusoidal [14]. The highest point of a human's trajectory in one gait cycle is the moment when the feet cross. This fact is deployed to implement the heel strike detection system.

2.1 Gait Trajectory Model Construction

The gait trajectory model [10] is described in Eq. (1). The modified version of this model is used and assumes that a subject walks at the same speed. This model only considers the vertical position of the object. The gait trajectory can be divided into two parts: a periodic factor and a scaling factor. The periodic factor is proportional to walking position; the scaling factor depends on imaging geometry. Therefore, the gait trajectory model is defined in the following way, where the vertical position y is a function of gait frequency ω.

First, the general case is

$$y = f(t) * \sin(\omega t + \theta) + g(t) \tag{1}$$

$$\text{periodic factor: } f(x) = C_1 \ln|\alpha(1 - t/\lambda)| + C_2 \tag{2}$$

$$\text{scaling factor: } g(x) = C_3 \ln|\alpha(1 - t/\lambda)| + C_4 + I_{gait} \tag{3}$$

where ω is a gait frequency, θ is the initial phase, I_{gait} is the initial position of gait, λ is a total time, α is a total length of walking, and C_n are constants.

Around the image center, the scaling factor can be modeled as

$$\text{scaling factor: } g_c(t) = C_0 t + I_{gait} \tag{4}$$

where C_0 is constant

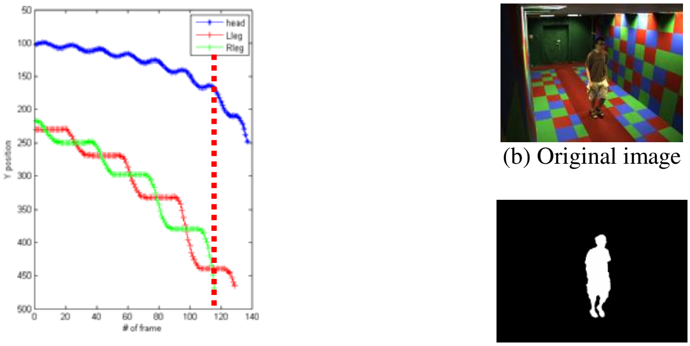
To clarify the model the exact human trajectory using some manually chosen points are extracted. We use 34 samples (24 males, 9 female, with around 40 images in each sequence) chosen at random from a gait biometric database [11] and extract the corresponding points for all frames. The extracted gait trajectories are normalized in order to express the error as a percentage. After fitting using the Levenberg-Marquadt algorithm R-squared and Sum of Square Error (SSE) is applied. Table 1 shows the numerical result of model fitting. The value of R-squared for all samples is over 98% and the sum of squared errors is less than 4%, both reflecting a good fit.

Table 1. The numerical data of each model fitting

Measure	Value (%)
SSE	3.87
R-Squared	98.0

2.2 Key Frame Selection

The key frame can be calculated after construction of the model. The highest points of trajectory are calculated from gait frequency f. ($x = ((2n+3/2)\pi - \theta)/2\pi f$, where n is any integer). Figure 1 shows a sample of the key frame extraction process. In fig. 1(a) the highest position of y is when the left foot and the right foot cross (here, the key frame number is 97). Fig. 1(b) and (c) is the original and silhouette image at the key frame. As shown in fig. 1(b) it indicates that the head is in the highest position is when the feet cross. In the next stage, the silhouette image is used for filtering the accumulator map to extract the ROI since the accumulator map is constructed using all of the images in a sequence and the filtered accumulator map must contain at least one heel strike.



(b) Original image

(a) Trajectories of y position for left and right feet and the head

(c) Silhouette image

Fig. 1. Key frame extraction

3 Heel Strike Detection

This section shows the process of detecting heel strike position using the pre-calculated key frame information. An accumulator map is used to be derived by adding samples of the walking subject's silhouette to determine which parts of the body remained longest at same position. Generally, during the strike phase, the foot of the striking leg stays at the same position for half a gait cycle, whilst the rest of the human body moves forward.

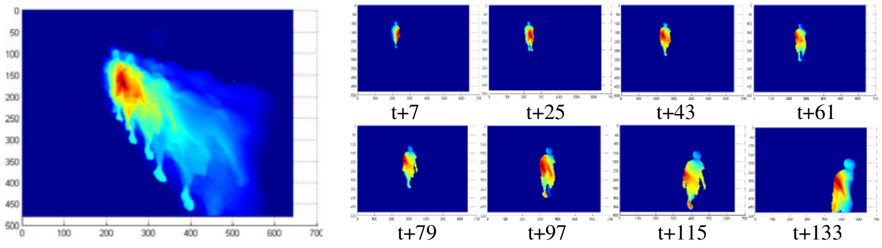
3.1 Heel Strike Candidate Extraction

As a preprocessing step, we calculate the silhouette image [12] from the intensity and the color difference (between the background image and foreground image) at each pixel. Then, the accumulator map of a silhouette is the number of silhouette pixels in $(i,j)^{th}$ position. Low pass filtering is deployed to smooth the accumulator surface.

$$Accumulator(i,j) = \sum^{\#\ of\ \ images} Silhouette\quad(i,j) \tag{5}$$

Figure 2 shows an accumulator map and filtering result of a key frame silhouette image. The colour in the figure indicates the number of silhouette pixels from blue (few) to red (many). As shown in fig. 2(a), the heel strike region can clearly be distinguished from the other body parts.

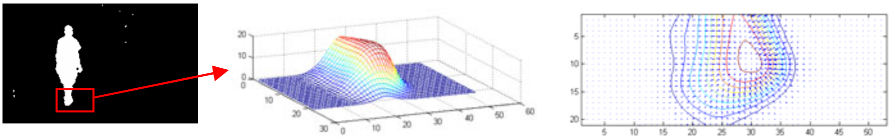
The filtered accumulator map shows the distribution of the number of silhouette pixels. It reveals that the position of heel strike has a relatively higher distribution than other regions. Using the characteristic, we extract a Region of Interest (ROI) and make it smoother to apply Gradient descent algorithm. Figure 3 shows the process for finding the heel strike positions from the accumulator map. The accumulator map shown in Fig. 2(a) is filtered by Gaussian function (filter size 12×12, $\sigma = 2.0$). Then, the approximate heel region, which is one eighth of person's silhouette height from the bottom of silhouette, is extracted (fig. 3(a)). Accordingly, the heel strike position can be extracted by gradient descent. Fig. 3(b) shows the three dimensional view of the extracted ROI. Fig. 3(c) shows the result of analyzing Fig. 3(b) using the Gradient Descent algorithm. The small arrow in the figure is the point where the orientation has changed. Fig. 3(c) shows the trace convergence to the local maximum.



(a) The accmulator map

(b) ROI extraction using filtering

Fig. 2. Silhouette accumulator map



(a) ROI extraction

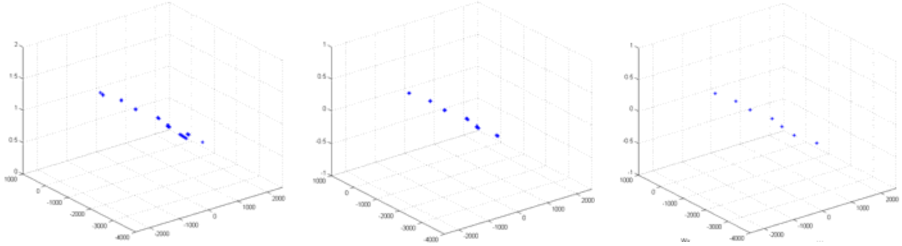
(b) 3D view of ROI

(c) Result of Gradient descent

Fig. 3. The procedure of heel strike detection

3.2 Heel Strike Position Verification

In previous section, the process for extracting the heel strike candidates is described. This section describes the heel strike verification process. In our method the silhouette image is used when the feet cross, so it is possible to extract the candidates from foot of heel strike and also another foot. For instance, in the second heel strike of fig. 4(a), the candidates are detected from both feet. To reduce the invalid candidates, the key



(a) The candidates

(b) The result of filtering using other candidates

(c) The filtering result using 3D position

Fig. 4. The verification process

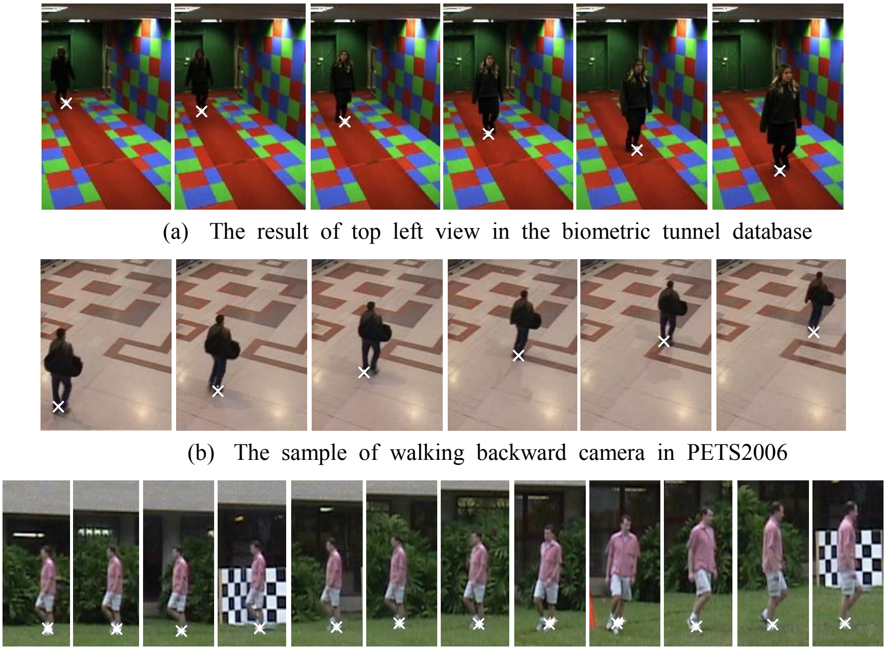
frame is calculated when the position of *y* is lowest in one gait cycle (fig. 1(a)) and the same procedure is executed in Section 3.1 to find other heel strike candidates. Since the moment at the lowest *y* is that the gait stride is largest, the feet still stay on the floor, and positions of the feet are separated, so the candidates from other key frames are considered as the potential heel strikes. These candidates are deployed to remove the invalid candidates. Simply, the distance between these two groups of candidates (at the highest and lowest *y* coordinate) is calculated. Then, the candidates in the fixed distance (here, 5 pixels) are selected from the group of candidates of lowest values for *y*. As shown in fig. 4(b), after this filtering process the invalid candidates from another foot are removed.

The accumulator map depends on camera view and once the camera is calibrated the invalid candidates could be removed using the back projection from a 2D image plane into a 3D world space. Using the 3D projection the candidates which are the closest from the camera are selected. Since a single camera is used in our approach, we assume a ground floor is known, i.e. that z=0 (the z axis is vertical position). This enables calculation of the intersecting points between the projection ray from 2D image points and the ground floor. The closest heel strike to the floor is considered as the final heel strike position, thereby filtering the invalid positions. Figure 4 shows a result of the filtering process and the invalid points in fig. 4(a) are removed to give the final result in fig. 4(c).

4 Experimental Results

To evaluate the proposal heel detection system, we use two visual surveillance databases: a biometric gait database [11] and PETS 2006 database [13]. The biometric tunnel data consists of 25 samples (18 males, 7 female, with around 130 images in each sequence) and each sample has two views of image sequences which are different from the data used to verify the gait trajectory model in Section 2. Moreover, we choose 10 samples from an image sequence of PETS2006 dataset which is recorded at a train station. In the dataset, each sample has a different walking direction with around 80 images in each sample sequence.

Figure 5(a) and (b) show the detection result with different environments; the biometric tunnel and a train station from the PETS data. The white crosses in the figure represent the points detected as the heel strike positions.



(a) The result of top left view in the biometric tunnel database

(b) The sample of walking backward camera in PETS2006

(c) The sample of walking direction change in GaitChallenge dataset

Fig. 5. Detection result with different environments

As shown in fig. 5 the proposed method can detect the heel strike position regardless the camera view since the method uses the basic characteristic of gait: its periodic factor. In the cases of above samples the subjects walk in series of straight lines. To confirm that our method works in the case of walking direction change a single image sequence is tested in Gait Challenge database [15] where the subject is walking around an elliptical path. In this situation, Equation 1 is not suited to finding the gait frequency. So, the gait frequency in the frequency domain is calculated since the waling speed is almost constant. Then, the same procedure is applied. Figure 5(c) shows the result of detection. Even the walking direction changes the proposed method still follows the position of heel strike.

Table 2 shows the result of detection rate. A total of 359 heel strikes were tested. We calculate the detection rate manually because the database does not provide ground truth of heel strike position. The detection rate for the biometric tunnel database is slightly higher than for the PETS 2006 database since the environment of the biometric tunnel is more controlled. The overall detection rate is 95.3%.

Table 2. The result of heel strike detection

Database	Value (%)
Biometric tunnel	95.6 (285/298)
PETS 2006	93.4 (57/61)

5 Conclusions

To deploy automatic gait recognition in unconstrained environments, we need to develop new techniques for analysis. This paper describes new techniques for heel strike estimation to be robust to feet self-occlusion and view of camera. The approach to heel strike estimation combines human walking analysis with characteristics of heel strike. The approach has been demonstrated on a visual surveillance database and on one for biometrics with a heel strike detection rate was over 95%. As such heel strike analysis can be used for basic gait analysis and derivation of walking direction estimation and this approach provides a new and more generalized approach for surveillance environments.

References

1. Bouchrika, I., Nixon, M.S.: Model-based feature extraction for gait analysis and recognition. In: Gagalowicz, A., Philips, W. (eds.) MIRAGE 2007. LNCS, vol. 4418, pp. 150–160. Springer, Heidelberg (2007)
2. Bobick, A.F., Johnson, A.Y.: Gait recognition using static, activity-specific parameters. In: Proc. CVPR, pp. 423–430 (2001)
3. BenAbdelkader, C., Cutler, R., Davis, L.: View-invariant estimation of height and stride for gait recognition. In: Tistarelli, M., Bigun, J., Jain, A.K. (eds.) ECCV 2002. LNCS, vol. 2359, pp. 155–167. Springer, Heidelberg (2002)
4. Jean, F., Bergevin, R., Albu, A.B.: Body tracking in human walk from monocular video sequences. In: Proc. of CRV, pp. 144–151 (2005)
5. Vignola, J., Lalonde, J.F., Bergevin, R.: Progressive human skeleton fitting. In: Proc. of ICVI, pp. 35–42 (2003)
6. Zhou, Z., Prugel-Bennett, A., Damper, R.I.: A Bayesian framework for extracting human gait using strong prior knowledge. IEEE TPAMI 28(11), 1738–1752 (2006)
7. Sigal, L., Black, M.J.: Measure locally, reason globally: occlusion-sensitive articulated pose estimation. In: Proc. CVPR, pp. 2041–2048 (2006)
8. Zhang, J., Luo, J., Collins, R., Liu, Y.: Body localization in still images using hierarchical models and hybrid search. In: Proc. CVPR, pp. 1536–1543 (2006)
9. Sundaresan, A., Chellappa, R.: Model-driven segmentation of articulating humans in laplacianeigenspace. IEEE TPAMI 30(10), 1771–1785 (2008)
10. Jung, S.U., Nixon, S.M.: On using gait biometrics to enhance face pose estimation. In: Proc. of IEEE BTAS, p. 6 (2010)
11. Seely, R.D., Samangooei, S., Middleton, L., Carter, J.N., Nixon, M.S.: The university of southampton multi-biometric tunnel and introducing a novel 3D gait dataset. In: Proc. of IEEE BTAS, p. 6 (2008)
12. Cheung, G., Kanade, T., Bouquet, J., Holler, M.: A real time system for robust 3d voxel reconstruction of human motions. In: Proc. CVPR, pp. 714–720 (2000)
13. PETS: Performance Evaluation of Tracking and Surveillance, `https://www.cvg.cs.rdg.ac.uk/slides/pets.html`
14. Aristotle: On the Motion of Animals, B.C. 350
15. Phillips, P.J., Sarkar, S., Robledo, I., Grother, P., Bowyer, K.W.: The gait identification challenge problem: data sets and baseline algorithm. In: Proc. ICPR, pp. 385–388 (2002)

Detecting Customers' Buying Events on a Real-Life Database

Mirela C. Popa[1], Tommaso Gritti[2], Leon J.M. Rothkrantz[1],
Caifeng Shan[2], and Pascal Wiggers[1]

[1] Man-Machine Interaction Group, Delft University of Technology,
Delft, The Netherlands
m.c.popa@tudelft.nl

[2] Video and Image Processing Group, Philips Research, Eindhoven, The Netherlands
{tommaso.gritti,caifeng.shan}@philips.com

Abstract. Video Analytics covers a large set of methodologies which aim at automatically extracting information from video material. In the context of retail, the possibility to effortlessly gather statistics on customer shopping behavior is very attractive. In this work, we focus on the task of automatic classification of customer behavior, with the objecting to recognize buying events. The experiments are performed on several hours of video collected in a supermarket. Given the vast effort of the research community on the task of tracking, we assume the existence of a video tracking system capable of producing a trajectory for every individual, and currently manually annotate the input videos with trajectories. From the annotated video recordings, we extract features related to the spatio-temporal behavior of the trajectory, and to the user movement, and analyze the shopping sequences using a Hidden Markov Model (HMM). First results show that it is possible to discriminate between buying and non-buying behavior with an accuracy of 74%.

Keywords: Trajectory analysis, Optical flow, Hidden Markov Models, Shopping Behavior.

1 Introduction

There is an increasing amount of research in the area of video analytics and semantic interpretation as an application to automatic surveillance, traffic monitoring, video games, marketing, etc. In the field of marketing it is of primary concern to identify the most appealing products and services for customers and to maximize their impact on the shopping behavior. Computer vision provides multiple techniques which enable surveillance [5], action recognition, and behavior interpretation of customers. Tracking people inside the shop can have many applications, such as global shopping behavior recognition, region of interest detection both individually and for a group of customers, measured at a specific moment or over time intervals. We plan to use the existing surveillance systems to observe the shopping behavior of people [4], to get a better understanding of

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 17–25, 2011.

their needs. The action recognition module can provide cues regarding customers' interest in products and can help interpreting different interaction patterns, such as grasping a product immediately, after a period time or even after more visits at the same place. In this paper we propose an automatic surveillance system for detecting customers' buying behavior based on tracking and motion information and tested on real-life recordings in a shopping mall. Its applicability resides in identifying different buying patterns in terms of number of interactions and time spent in the vicinity of a product but also in finding for which products categories the customers have trouble deciding. As a result, appropriate actions could be taken such as new products arrangements and more efficient usage of the store space. Next we provide an overview of related studies, then the design of our system is presented, followed by the data acquisition process and the experimental results section. Finally we formulate our conclusions and give directions for future work.

1.1 Related Work

People tracking, behavior analysis, and prediction were investigated by Kanda et al. in [3]. Accumulated people's trajectories over a long period of time provided a temporal use-of-space analysis facilitating the behavior prediction task performed by a robot. Hu et al. [2] used the Motion History Image (MHI) along with the foreground image obtained by background subtraction and the histogram of oriented gradients (HOG) [1] to obtain discriminative features for action recognition. Next a multiple-instance learning framework SMILE-SVM was build to improve the performance. This approach proved its effectiveness on a real world scenario from a surveillance system in a shopping mall aimed at recognizing customers' interest in products defined by the intent of getting the merchandize from the shelf. These approaches are suitable for action recognition under varying conditions in complex scenes such as background clutter or partial occluded crowds; still they require supervised learning based on a large reliable dataset. Human behavior analysis while shopping was investigated by Sicre and Nicolas in [7]. They propose a finite-state-machine model for simple actions detection, while the interaction between customers and products is based on MHI and accumulated motion image (AMI) [8] description and SVM classification. It remains to be proved and tested whether this method will be applicable in an uncontrolled real-life scenario which deals with occlusions and different types of settings. Another issue regards the variability of performing an action in relation with the dataset size, which in this case is limited to 4 persons.

2 Proposed Methodology

Based on observations made in real shops we proposed a number of shopping behavior models as described in [4]. There are many individual differences in shopping behavior of people. Some shoppers know what they want and the location of that product (*goal oriented*), others prefer to inspect the offer (*looking* *around*), some are helpless and would need support (*disoriented*), while others are actively looking for assistance, finally some shoppers are just looking for interesting products or just enjoy being in a shop (*fun-shopper*). We assume the ultimate goal of shopping is to buy a required product. Next the design of our system for automatic assessment of customers' buying behavior is presented. We propose a modular approach and we describe next the functionality of each module. A diagram of the proposed system is shown in Fig. 1.

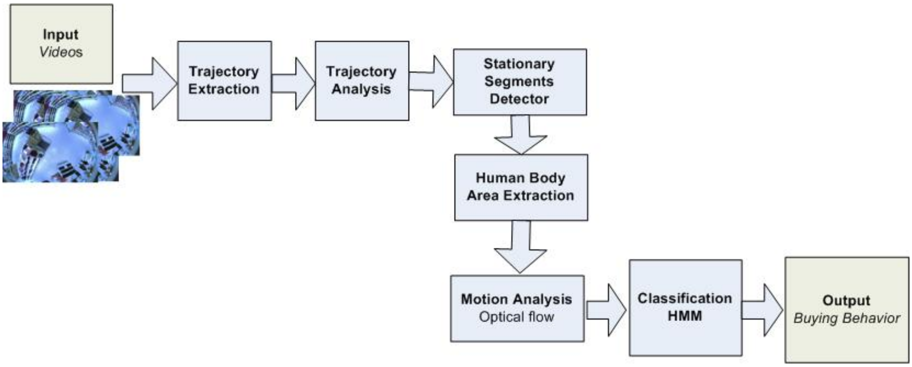


Fig. 1. System Overview

2.1 Trajectory Extraction

First the trajectory extraction module is employed. Currently the customers' trajectories are manually labeled, given that our goal consisted in the high-level analysis of behavior. In our future work, trajectories will be extracted by adopting person detection and tracking. For this task we used our frame based annotation tool which enables both person and event annotation.

2.2 Trajectory Analysis

Global motion analysis provides a first insight into customers' shopping behavior. Therefore the Trajectory Analysis module is employed to extract relevant trajectory features. We started from the feature set $f\mathrm{T} = [x, y, x^{'}, y^{'}, x^{''}, y^{''}]$ proposed in [11], described by position (x,y), velocity $(x^{'}, y^{'})$, and acceleration $(x^{''}, y^{''})$. We decided to exclude spatial features (x,y) in order to prevent learning of a preferred shopping path, while our interest resided in capturing motion characteristics. The curvature k of a trajectory was considered due its properties such as invariance under planar rotation and translation of the curve [9].

$$k = (y^{''}/x^{'})^2/(\sqrt{(}(y^{'}/x^{'})^2 + 1))^3$$

Based on experiments we noticed that the best feature set for encapsulating trajectory information was the following one: $f\mathrm{T} = [x^{'}, y^{'}, x^{''}, y^{''}, \sqrt{(}\Delta x^2 + \Delta y^2), k]$, where $\Delta x = x(t) - x(t-1)$ and $\Delta y = y(t) - y(t-1)$.

2.3 Stationary Segments Detector

The next module of our system is responsible for detecting segments of interest and potential buying segments. The detection is performed using the features defined in the previous section. Due to non-linearity of persons' motion and errors introduced by the manual annotation, we used Gaussian smoothing of velocity. In this way each velocity value v_μ is approximated by:

$$v_\mu = \sum_{t=\mu-\sigma}^{t=\mu+\sigma} v_t * N(t;\mu,\sigma^2) \quad (2)$$

where N is a density function for normal distribution with the mean μ and variance σ^2.

2.4 Human Body Area Extraction

For each trajectory segment detected by the previous module, the human body area is extracted. To this aim, for every frame in a segment, we estimate a binary mask corresponding to a human in a given trajectory point, according to [6]. We then combine all binary masks belonging to a segment into one area. The combined binary mask is used to extract image content from every frame. The extracted image content is rectified along the radial direction (see an example in Fig. 2), to remove the influence of the orientation, i.e. so that all people are in the upright position.

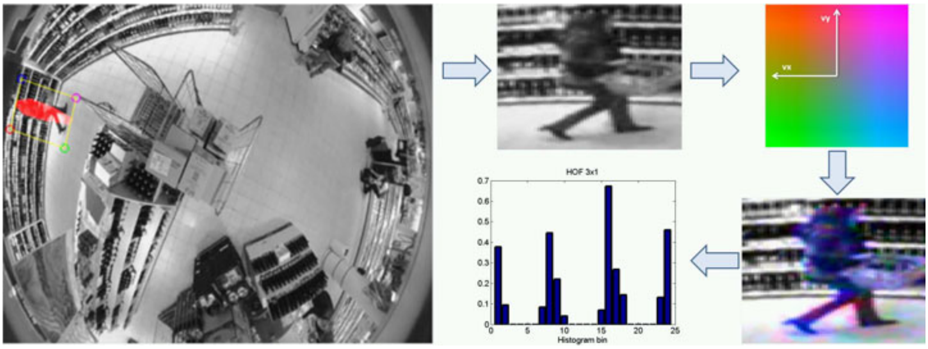


Fig. 2. Overview of the Human Body Area Extraction and Motion Analysis modules. From left to right, clockwise: human binary mask from [6], highlighted in red; rectified area defined by the combination of binary mask in the stationary segment; optical flow and corresponding color coding; histogram of optical flow.

2.5 Motion Analysis

We assume buying behavior can be characterized by motion patterns, such as picking a product and putting it in the shopping basket. The motion analysis module is applied to each segment, by estimating optical flow in the rectified

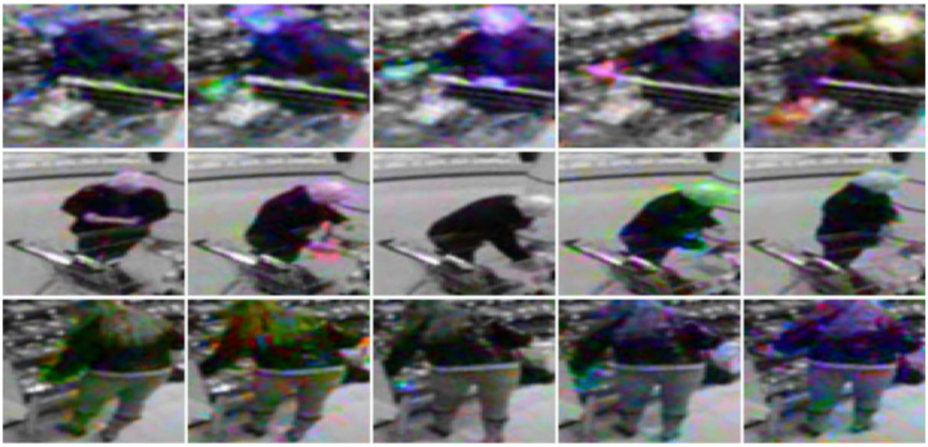


Fig. 3. Motion Flow Visualization. Overlay color is according to color coding shown in Fig. 2.

areas between every two consecutive frames. Normalized histograms of motion vectors in 8 directions are extracted from the whole image patch and also by considering a image patch segmentation into three regions corresponding to the approximate position of the head, body and legs of a person. We tested several optical flow algorithms both in terms of accuracy and also execution time such as Lucas-Kanade or Horn-Schunk and the best results were obtained using the method proposed by Liu [10]. An example of a buying event is depicted in Fig. 3.

2.6 Classification

Classification techniques can be divided into two groups, namely supervised and unsupervised. From the supervised group (e.g. Hidden Markov Models (HMMs), SVM, and Gaussian Mixture Models (GMMs)) we chose a HMM-based classification method due to its characteristics such as incorporating dynamics of motion features during time and ability to capture temporal correlations. The extracted features (trajectory and optical flow) were fed to a HMM and the maximum likelihood rule was used to decide the label corresponding to each interesting trajectory segment.

3 Experimental Results

3.1 Data Acquisition

In order to test our system in a realistic environment, we recorded video material in a supermarket, at different time intervals, using a fish-eye camera attached to the ceiling. An example of the acquired type of images is shown in Fig. 2.

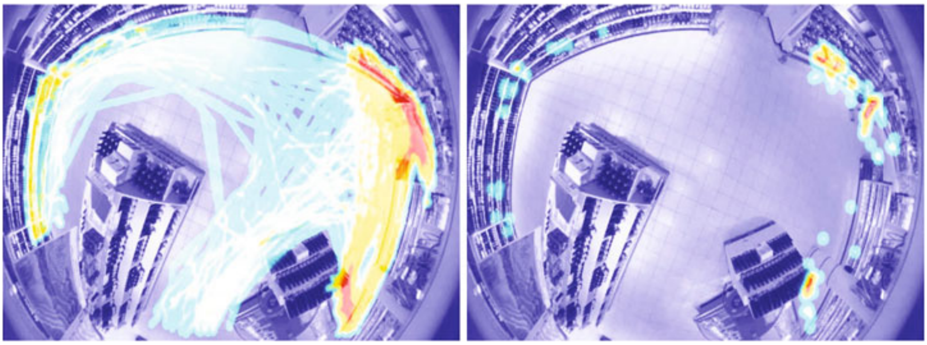


Fig. 4. Trajectory density map (left), and buying event density (right), computed on the dataset adopted for the experimental evaluation. Color coding shows in red areas with higher density.

We collected and manually annotated approximately 5 hours of recordings resulting in 270 customers' trajectories, from which 100 trajectories contained buying segments. We define as a *buying event* an action of a customer who picked a product and put it in the shopping basket or just took it with him/her. The total number of annotated buying events was 130, since some of the trajectories contained more than one buying event. A density map of the annotated trajectories and for the buying events is shown in Fig. 4. We present next the experimental results obtained using the recorded data.

3.2 Experiments

We performed a number of tests in order to find the best feature descriptor and HMM topology for our buying behavior analysis system as described in Section 2. We investigated different detection methods of potential buying trajectory segments. The aim was to detect automatically the segments containing buying events. From our observations of buying behavior we noticed that the action of buying a product usually happens after the customer stopped for a period of time in the products area. By employing a stationary detector as described in Section 2.3 based on slow velocity and duration of staying in the vicinity of a product of at least one second we were able to detect 90% of all the buying actions, meaning 118 segments out of 130. The rest of 10% were associated with a different type of behavior (goal oriented) characterized by a customer which knows what he wants and picks that product very quickly and then continues his shopping trip. By applying the stationary detector the number of analyzed video frames (N) was reduced to 67% of which 17% corresponded to buying segments and 50% to non-buying ones.

In order to refine our analysis, we employed motion analysis in the detected stationary trajectory segments. Normalized histograms of optical flow (HOF) was selected as feature. Adopting a quantization of the optical flow directions in 8 bins proved to be the best tradeoff compared to average length and angle. Furthermore, we investigated the influence of computing optical flow histogram in separate regions of the rectified image patch, and concatenating them to allow for an increased level of detail. We refer to HOF for the case of a single histogram, HOF3x1 for the case of subdivision of the image in 3 vertical subregions, and HOF3x3 for the subdivision of 9 subregions. The performance of a HMM is highly dependant on its topology. In order to determine the best topologies for our HMM models we performed an extensive search, by employing a diverse number of states (1-10), number of Gaussian Mixtures (1-20), and also network topologies (left-to-right, ergodic model). We found out that the best accuracy of 74% was obtained for a HMM model (left-to-right) with 6 states and 2 GMMs, for HOF3x3, using a 9-fold cross validation testing approach. The ROC curves obtained for the different HOF features are shown in Fig. 5. The improvement in accuracy of HOF3x3 over HOF3x1 and HOF features indicates that such separation allows to better discriminate actions, possibly because of different body parts movement which are related to the buying actions.

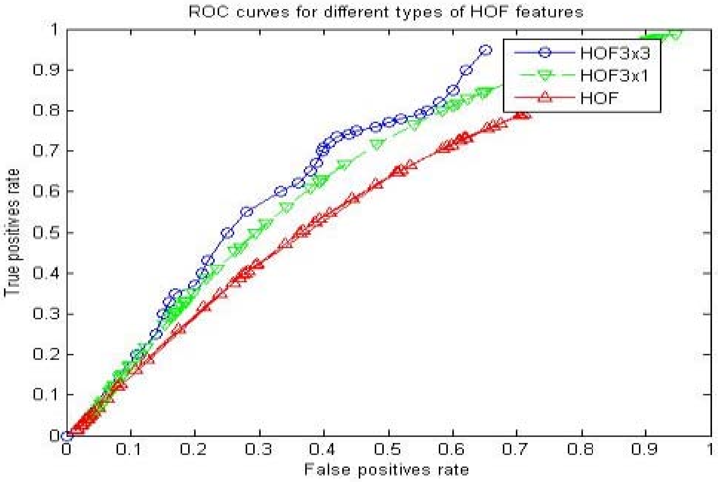


Fig. 5. ROC curves for HOF features

The number of HMM training iterations plays an important role at determining the best tradeoff between true positives and false positives rate. In our case 10 iterations were enough to learn the two HMM models, while using a bigger number of iterations led to overfitting the non-buying HMM model in detriment of the buying one. The miss-classified false negatives are due to the challenging type of data, such as occluded buying sequences either by another customer or by the person herself, while the false positives contain examples of customers' picking a product and then putting it back or just interacting with the shopping basket without putting a new product. Still given the difficulty and variability of the recorded data we consider that our approach is quite good at detecting customers' buying behavior under varying conditions.

4 Conclusion and Future Work

We presented an approach towards understanding customers' shopping behavior applied to real-life recordings in a supermarket. We designed and implemented a first running prototype for detecting customers' buying behavior. We used global features extracted from trajectories to detect potential buying segments and we extracted optical flow from an interesting area for action recognition. We achieved a best accuracy of 74% by using HOF3x3 features. As future work we plan to improve and refine the action recognition module by using different types of features such as interest-points models and an extended set of shopping related actions. We also aim at extending the system by fusing the data from cameras at different location and view angles.

Acknowledgement. This work was supported by the Netherlands Organization for Scientific Research (NWO) under Grant 018.003.017 and the Visual Context Modeling (ViCoMo) project.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2005), San Diego, California, vol. 1, pp. 886–893 (June 2005)
2. Hu, Y., Cao, L., Lv, F., Yan, S., Gong, Y., Huang, T.S.: Action detection in complex scenes with spatial and temporal ambiguities. In: Proceedings of International Conference on Computer Vision, ICCV 2009 (October 2009)
3. Kanda, T., Glas, D.F., Shiomi, M., Ishiguro, H., Hagita, N.: Who will be the customer? A social robot that anticipates people's behavior from their trajectories. In: Int. Conf. on Ubiquitous Computing, UbiComp 2008 (2008)
4. Popa, M.C., Rothkrantz, L.J.M., Yang, Z., Wiggers, P., Braspenning, R., Shan, C.: Analysis of Shopping Behavior based on Surveillance System. In: 2010 IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC 2010), Istanbul, Turkey (2010)
5. Valera, A., Velastin, S.A.: Intelligent distributed surveillance systems: A Review. IEEE Proc. Vision, Image, and Signal Processing 152(2), 192–204 (2005)
6. Zhang, Z., Venetianer, P.L., Litpon, A.J.: A Robust Human Detection and Tracking System Using a Human-Model-Based Camera Calibration. In: The Eighth International Workshop on Visual Surveillance (2008)
7. Sicre, R., Nicolas, H.: Human behavior recognition at a point of sale. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Chung, R., Hammound, R., Hussain, M., Kar-Han, T., Crawfis, R., Thalmann, D., Kao, D., Avila, L. (eds.) ISVC 2010. LNCS, vol. 6455, pp. 635–644. Springer, Heidelberg (2010)
8. Kim, W., Lee, J., Kim, M., Oh, D., Kim, C.: Human action recognition using ordinal measure of accumulated motion. EURASIP Journal on Advances in Signal Processing (2010)

9. Junejo, I.N., Javed, O., Shah, M.: Multi Feature Path Modeling for Video Surveillance. In: 17th Int. Conf. on Pattern Recognition (ICPR 2004), vol. 2 (2004)
10. Liu, C.: Beyond Pixels: Exploring New Representations and Applications for Motion Analysis, Doctoral Thesis. Massachusetts Institute of Technology (May 2009)
11. Moris, B.T., Trivedi, M.M.: A survey of vision-based trajectory learning and analysis for surveillance. IEEE Trans. on Circuits and Systems for Video Technology 18(8), 1114–1127 (2008)

A Simplified Gravitational Model for Texture Analysis

Jarbas J. de M. Sá Junior[1] and André R. Backes[2,*]

[1] Departamento de Engenharia de Teleinformática - DETI
Centro de Tecnologia - UFC
Campus do Pici, S/N, Bloco 725
Caixa Postal 6007, CEP: 60.455-970, Fortaleza, Brasil
jarbas_joaci@yahoo.com.br

[2] Faculdade de Computação
Universidade Federal de Uberlândia
Av. João Naves de Ávila, 2121
38408-100, Uberlândia, MG, Brasil
backes@facom.ufu.br

Abstract. Textures are among the most important features in image analysis. This paper presents a novel methodology to extract information from them, converting an image into a simplified dynamical system in gravitational collapse whose states are described by using the lacunarity method. The paper compares the proposed approach to other classical methods using Brodatz's textures as benchmark.

Keywords: Texture Analysis, Simplified Gravitational System, Complexity, Lacunarity.

1 Introduction

Texture analysis is one of the most important fields in computer vision. Although there is no formal definition about the concept of texture, it is easily identified by humans and it is rich in visual information. In general, textures are complex visual patterns composed by entities, or sub-patterns, with bright, color, orientation and size characteristics [1]. So, textures supply very useful informations for automatic recognition and interpretation of an image by a computer [2].

Over the years, many methods of texture analysis have been developed, each of them obtaining information in a different way. Actually, most of the methods can be grouped into four main categories [3]: statistical, geometrical, model-based, signal processing methods. The statistical approach includes classical methods, such as co-occurrence matrices [1]. An important example of signal processing methods is the Gabor filters [4]. Still in this category, we find many studies on texture analysis in spectral domain, especially after the invention of wavelet transform (e.g., [5,6]).

[*] Is grateful to CNPq (Grant #453371/2011-6).

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 26–33, 2011.

In recent years, many other approaches have been developed to study pixels' relationship. One of these aimed to represent and characterize the relation among pixels using the Complex Networks Theory [7]. Another important approach is the Tourist Walk [8], where each pixel is a tourist wishing to visit N cities according to the following rule: go to the nearest city that has not been visited in the last μ time steps. Fractal analysis has recently been proposed as a replacement to the histogram to study the distributions of gray levels along a texture pattern as well [9].

In order to explore images in a new manner, and, therefore, extract valuable information from them, this work presents a novel approach, which transforms an image in a dynamic system in gravitational collapse process. This approach enables images to evolve and to present different states, each of which offering a new source of information to be extracted. To accomplish this, we employed the lacunarity method to quantify each state in order to obtain a feature vector.

Our presentation is composed as follows. Section 2 shows the rules established to simulate a simplified gravitational collapse in an image. Section 3 describes the process of composing image signatures by applying the lacunarity method in states of the collapse process. In Section 4, we described an experiment that uses 40 classes of Brodatz's textures (10 images per class). Section 4.1 demonstrates the superior performance of the proposed approach when it is compared to results of other important methods. Finally, we made some considerations of this work in Section 5.

2 Texture Analysis and Simplified Gravitational System

When an object orbits another one, two forces need to be considered. The first is the gravitational force, which was stated by Isaac Newton in "the Principia" and can be defined by the following sentence: the force exerted by an object to another one is directly proportional to the product of their masses and inversely proportional to the square of the distance between them [10]. This force, as illustrated by Figure 1, is given by the following equation

$$\boldsymbol{f}_a = \frac{G.m_1.m_2}{\|\boldsymbol{r}\|^2} . \frac{\boldsymbol{r}}{\|\boldsymbol{r}\|} \qquad (1)$$

where G is the gravitational constant, m_1 and m_2 are the masses of the two particles, $\boldsymbol{r}$ is the vector connecting the positions of the particles and $f_{1,2}$ is the gravitational force between the two particles. $\|\|$ denotes the magnitude or norm of a vector.

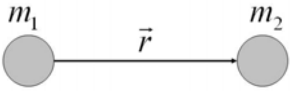


Fig. 1. Example of gravitational force between two massive particles

The other force is the centripetal force, which is directed to the center of a circular trajectory described by an object and is directly proportional to its tangential speed. This force can be defined by the following equation

$$\boldsymbol{f}_c = m\boldsymbol{a}_c = m\frac{\boldsymbol{v}^2}{\|\boldsymbol{r}\|} \tag{2}$$

where f_c is the centripetal force, m is the mass, $\boldsymbol{a}_c$ is the centripetal acceleration, $\boldsymbol{v}$ is the tangential speed and $\|\boldsymbol{r}\|$ is the radius of a circular trajectory.

Before applying these concepts on a texture image, some considerations are necessary. Literature commonly describes a gray-scale texture as a bi-dimensional structure of pixels. So, let $I(x, y) = 0 \ldots 255$, $(x, y = 1 \ldots N)$, be a pixel in an image I, where x and y are the Cartesian coordinates of the pixel. The integer values associated to a pixel $I(x, y)$ represents the gray-scale of that pixel.

We considered each image pixel $I(x, y)$ as a particle in the gravitational system, where the intensity associated to that pixel, $I(x, y)$, is its mass m. In a real gravitational system, all particles interact with each other. Instead of adopting this approach, which has a high computational cost, we considered that there is just interaction between each pixel and an object of mass M, located at the center of the texture image.

For each pixel, we established a gravitational force according to Equation 1, where we replaced mass m_1 by M and m_2 by the gray-scale of the pixel, and a centripetal force according to Equation 2, where the pixel intensity replaces m and determines the tangential speed. To determine this speed, we have to take into account that a very low speed causes a very fast collapse and, therefore, information loss, while high speeds imply no collapse.

So, in order to find a range of tangential speeds so that all pixels could collapse slowly, we established $f_a = f_c$, thus yielding the highest tangential speed as described by the following function

$$\boldsymbol{v}_{max} = \sqrt{\frac{GM}{r_{max}}} \tag{3}$$

where $\boldsymbol{v}_{max}$ is the highest speed of a pixel and r_{max} is the greatest distance between a pixel and image center. This speed assures that even the farthest pixel from the image center will collapse, that is, the pixel will gradually approaches the center of the image.

To extract information regarding to both distance and gray-level intensity, each pixel has its speed determined according to the following equation

$$\boldsymbol{v}_{pix} = \left(1 + \frac{I(x, y)}{255}\right)\frac{\boldsymbol{v}_{max}}{2} \tag{4}$$

where $\boldsymbol{v}_{pix}$ is the tangential speed of the pixel and $I(x, y)$ is its the gray-level. In this way, each pixel has a particular trajectory defined by its distance and its intensity, giving image its own signature.

These rules give pixels to types of movement. The first is constant, anti-clockwise circular, defined by $S_1 = \boldsymbol{v}_{pix}.t$, where D_1 is the distance covered by the pixel in a time t. To compute this new position the pixel is rotated using an angle of $S_1/(2\pi r)$, where r is the distance between the pixel and the center of the image.

The second movement is accelerated rectilinear, directed to the center of the image. The new location of the pixel is computed as $S_2 = (1/2).\boldsymbol{a}_{pix}.t^2$, where $\boldsymbol{a}_{pix}$ is the acceleration of the pixel toward the image center, given by $(\boldsymbol{f}_a - \boldsymbol{f}_c)/I(x,y)$ (for pixels $I(x,y) = 0$ (no mass), this equation becomes $0/0$, which we considered 0, that is, the pixels only rotate), and D_2 indicates the space covered by the pixel in a time t. To compute this new position, we decrement/increment the axis x and y using the proportion S_2/r. Figure 2 shows the movement of a pixel in a determined time t.

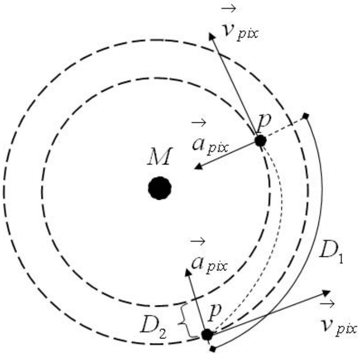


Fig. 2. Example of a simplified gravitational model where a pixel p collapses. The new position of the pixel is defined by the distances D_1 and D_2.

By applying this gravitational model to images, eventually two or more pixels may try to occupy the same position during the collapse process. If this situation occurs, the position will receive the average of pixels' gray-levels. This adaptation aimed to reduce the complexity of the method and to preserve image information.

3 Signature for Collapsing Texture Patterns

In this section, we present an approach to extract a texture signature using the proposed collapsing model and a traditional texture descriptor, the lacunarity. The concept of lacunarity was introduced by Mandelbrot [11] to characterize different texture patterns that presented the same fractal dimension. Initially proposed for binary patterns, the lacunarity describes the texture according to the number of gaps dispersed over it. It is considered as a multi-scaled measure of texture's heterogeneity, since the lacunarity measured depends on the gap size [12].

The gliding-box algorithm is often used to compute the lacunarity due to its simplicity [12,13]. The method consists of gliding a box of size r over the texture pattern and to count the number of gaps existent in the binary pattern. Over the years, this approach was also extended to deal with grayscale images [14,15]. Instead of simply counting the number of gaps, these approaches compute the minimum u and maximum v pixel values inside the box. According to these values, a column with more than one cubic may be needed to cover the image intensity coordinates. This relative height of the column is defined as $S = v - u - 1$. By considering each possible box position in the image, we compute the probability distribution $Q(S, r)$ of the relative height for a box size r. Then, the lacunarity is achieved as

$$\Lambda(r) = \frac{\sum S^2.Q(S,r)}{\left[\sum S.Q(S,r)\right]^2} \tag{5}$$

During the collapse of a texture pattern, its roughness changes. That means its lacunarity is different for each collapsing time steps t. Thus, this collapsing approach enables us to characterize a texture pattern through the variations of its lacunarity. Thus, we propose a feature vector that represents the texture pattern in different collapsing time steps t by a set of lacunarity values computed for a given box size r:

$$\boldsymbol{\psi}_{t_1,t_2,\ldots,t_M}(r) = \left[\Lambda_{t_1}(r), \Lambda_{t_2}(r), \ldots, \Lambda_{t_M}(r)\right]. \tag{6}$$

We must emphasize that the lacunarity is a multi-scaled measure, i.e., it depends on the box size r [12]. Thus, it is convenient to consider a second feature vector that exploits such characteristic. Therefore, we propose a second feature vector that analyses the collapsing texture using different lacunarity values. This is accomplished by the concatenation of the signatures calculated using $\boldsymbol{\psi}_{t_1,t_2,\ldots,t_M}(r)$, for different r values.

$$\boldsymbol{\varphi}(r_{max}) = \left[\boldsymbol{\psi}_{t_1,t_2,\ldots,t_M}(2), \ldots, \boldsymbol{\psi}_{t_1,t_2,\ldots,t_M}(r_{max})\right] \tag{7}$$

where r_{max} is the maximum box size allowed.

4 Experiment

In order to evaluate the proposed feature vectors, an experiment using a synthetic texture database was set. This database consists of a set of 400 texture images extracted from the book of Brodatz [16]. Each sample presents 200 × 200 pixels size, with 256 gray levels. The samples are grouped into 40 classes, with 10 samples each. Evaluation of the proposed feature vectors was performed using a statistical approach. For this task, we used the Linear Discriminant Analysis (LDA) in a leave-one-out cross-validation scheme [17].

To provide a more robust evaluation of the proposed method, we also included a comparison with traditional texture analysis methods. For this comparison, the following methods were considered: Fourier descriptors [18], Co-occurrence matrices [1], Gabor filters [19], Tourist Walk [8].

4.1 Results

To apply the method over the set of images previously described some parameters have to be set. These parameters are the mass M and the gravitational constant G. The mass M can be understood as a massive black hole at the center of the image. This mass should be capable of attracting farther and darker pixels of the image. To accomplish this task, its value was empirically established as $M = 500$. The gravitational constant was set as $G = 1$ in order to give pixels a suitable step during the collapsing process. Figure 3 shows an example of collapse process of a Brodatz's texture in three different time steps using the cited values of parameters.

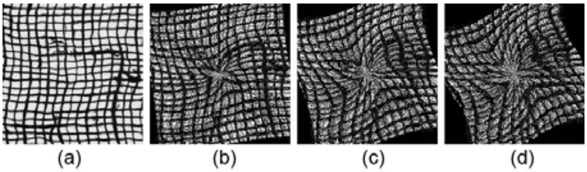


Fig. 3. Examples of collapsing texture images: (a) Original image; (b)-(d) Collapsing textures for time steps $t = \{10, 20, 30\}$

First, we analyzed the behavior of the estimated lacunarity using different box sizes as an image collapses (Figure 4). We note that different r values will lead to a different estimation of the lacunarity. However, the changes in the lacunarity value $\Lambda(r)$ are subtle as the collapsing time t increases, independent of the box size used. As a consequence, a feature vector $\boldsymbol{\psi}$ built using sequential time steps would present a large amount of redundant information and it should be avoided. The use of spaced values of t is preferred. It is important to remember that the lacunarity is considered as a multi-scaled measure. This is evident in Figure 4, where the lacunarity $\Lambda(r)$ is different for each box size r considered.

According to the previous considerations about the time step t and box size r, we propose to use both multiple r values and different sets of t values to compose the feature vector $\boldsymbol{\varphi}$. Then, this feature vector was used to characterize the Brodatz's samples in the proposed experiment. Table 1 presents the results achieved. The best result (97.00%) is obtained when $t = \{1, 6, 12, 18\}$ and $r_{max} = 11$ are used. Results indicate that the performance of the method increases as the number of time steps t and maximum box size r_{max} increase. However, for $r_{max} > 11$, a small decrease is perceived in the success rate. This indicates that larger box sizes are not efficient to capture the local characteristics of the texture pattern.

Table 2 presents the results obtained by each method compared. In this comparison, we considered the configuration that leads to the best results of our method in Table 1. Results demonstrate that our approach is a feasible texture descriptor as its results overcomes all traditional methods used during the comparison experiment.

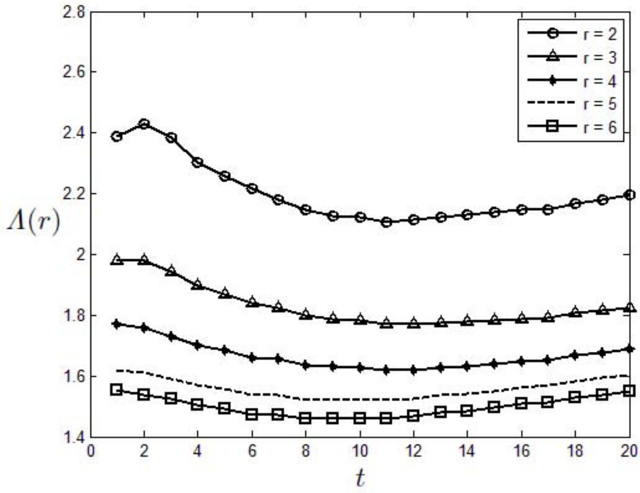


Fig. 4. Lacunarity estimated for time steps $t = 1, \ldots, 20$ and box sizes $r = \{2, 3, 4, 5, 6\}$

Table 1. Success rate (%) of the method on the Brodatz database for the φ feature vector

	r_{max}											
Time (t)	3	4	5	6	7	8	9	10	11	12	13	14
$\{1, 6\}$	84.50	89.50	90.25	90.75	91.50	92.50	92.75	93.25	94.00	93.50	93.75	93.25
$\{1, 6, 12\}$	89.75	92.50	93.75	94.00	93.50	94.75	95.00	94.75	96.75	96.75	96.25	95.75
$\{1, 6, 12, 18\}$	89.00	94.50	94.50	94.25	95.00	95.00	95.25	96.00	97.00	96.75	96.25	95.75

Table 2. Comparison results for different texture methods in the Brodatz database

Method	Images correctly classified	Success rate (%)
Fourier	351	87.75
Co-occurrence matrices	330	82.50
Gabor Filters	381	95.25
Tourist Walk	382	95.50
Proposed approach	388	97.00

5 Conclusion

This work presents a novel method to extract information from textures by transforming them in a simplified gravitational system whose states of collapse are explored by the use of lacunarity method. This unpublished approach showed results superior to the results yielded by classical methods, when tested on a Brodatz's database. Results showed that the method presents the best results by using spaced values of t and a set of r values. Thus, the proposed method opens a new research in texture analysis and amplifies the set of methods of identifying textures, improving the precision of the systems already developed.

References

1. Haralick, R.M.: Statistical and structural approaches to texture. Proc. IEEE 67(5), 786–804 (1979)
2. Bala, J.W.: Combining structural and statistical features in a machine learning technique for texture classification. In: IEA/AIE, vol. 1, pp. 175–183 (1990)
3. Tuceryan, M., Jain, A.K.: Texture analysis. In: Chen, C.H., Pau, L.F., Wang, P.S.P. (eds.) Handbook of Pattern Recognition and Computer Vision, pp. 235–276. World Scientific, Singapore (1993)
4. Casanova, D., Sá Junior, J.J.M., Bruno, O.M.: Plant leaf identification using gabor wavelets. International Journal of Imaging Systems and Technology 19(1), 236–243 (2009)
5. Lu, C.S., Chung, P.C., Chen, C.F.: Unsupervised texture segmentation via wavelet transform. Pattern Recognition 30(5), 729–742 (1997)
6. Arivazhagan, S., Ganesan, L.: Texture classification using wavelet transform. Pattern Recognition Letters 24(9-10), 1513–1521 (2003)
7. Costa, L.F., Rodrigues, F.A., Travieso, G., Villas Boas, P.R.: Characterization of complex networks: A survey of measurements. Advances in Physics 56(1) (2005)
8. Backes, A.R., Gonçalves, W.N., Martinez, A.S., Bruno, O.M.: Texture analysis and classification using deterministic tourist walk. Pattern Recognition 43, 685–694 (2010)
9. Varma, M., Garg, R.: Locally invariant fractal features for statistical texture classification. In: International Conference on Computer Vision - ICCV 2007, pp. 1–8 (2007)
10. Newton, I.: Philosophiae Naturalis Principia Mathematica. University of California, Berkeley (1999); original 1687, translation guided by I.B. Cohen
11. Mandelbrot, B.: The fractal geometry of nature. Freeman, San Francisco (1982)
12. Allain, C., Cloitre, M.: Characterizing the lacunarity of random and deterministic fractal sets. Phys. Rev. A 44(6), 3552–3558 (1991)
13. Facon, J., Menoti, D., de Albuquerque Araújo, A.: Lacunarity as a texture measure for address block segmentation. In: Sanfeliu, A., Cortés, M.L. (eds.) CIARP 2005. LNCS, vol. 3773, pp. 112–119. Springer, Heidelberg (2005)
14. Dong, P.: Test of a new lacunarity estimation method for image texture analysis. International Journal of Remote Sensing 21(17), 3369–3373 (2000)
15. Du, G., Yeo, T.S.: A novel lacunarity estimation method applied to SAR image segmentation. IEEE Trans. Geoscience and Remote Sensing 40(12), 2687–2691 (2002)
16. Brodatz, P.: Textures: A photographic album for artists and designers. Dover Publications, New York (1966)
17. Everitt, B.S., Dunn, G.: Applied Multivariate Analysis, 2nd edn. Arnold (2001)
18. Azencott, R., Wang, J.P., Younes, L.: Texture classification using windowed fourier filters. IEEE Trans. Pattern Anal. Mach. Intell. 19(2), 148–153 (1997)
19. Idrissa, M., Acheroy, M.: Texture classification using gabor filters. Pattern Recognition Letters 23(9), 1095–1102 (2002)

Robustness and Modularity of 2-Dimensional Size Functions – An Experimental Study*

Silvia Biasotti[1], Andrea Cerri[2], and Daniela Giorgi[1]

[1] IMATI, Consiglio Nazionale delle Ricerche, Genova, Italy
{silvia,daniela}@ge.imati.cnr.it
[2] Vienna University of Technology, Faculty of Informatics,
Pattern Recognition and Image Processing Group, Austria
acerri@prip.tuwien.ac.at

Abstract. This paper deals with the concepts of 2-dimensional size function and 2-dimensional matching distance. These are two ingredients of (2-dimensional) Size Theory, a geometrical/topological approach to shape analysis and comparison. 2-dimensional size functions are shape descriptors providing a signature of the shapes under study, while the 2-dimensional distance is the tool to compare them. The aim of the present paper is to validate, through some experiments on 3D-models, a computational framework recently introduced to deal with 2-dimensional Size Theory. We will show that the cited framework is modular and robust with respect to noise, non-rigid and non-metric-preserving shape transformations. The proposed framework allows us to improve the ability of 2-dimensional size functions in discriminating between shapes.

Keywords: multidimensional persistence, non-rigid shape analysis.

1 Introduction

Interpreting and comparing shapes are challenging issues in Computer Vision, Computer Graphics and Pattern Recognition [16,18]. Persistent Topology – including Size Theory [3] and Persistent Homology [11] – offers both theoretical and computational tools for shape comparison. The main idea is to take into account topological shape features with respect to some geometric properties conveyed by real functions defined on the shape itself [3].

Formally, this implies that a shape is represented by a pair (X, φ), where X is a topological space and $\varphi : X \to \mathbb{R}$ is a continuous real-valued function called *measuring function*. A number of descriptors have been introduced to describe pairs (X, φ) – such as Size Functions [12] – and successfully used for comparing images [7] and 3D models [4].

Nonetheless, a single real-valued measuring function φ may not be enough to cope with complex shape description problems. In fact, data are often characterized by two or more properties; this happens for example with physical simulations, where several measurements are made about an observed phenomenon,

* Partially supported by the CNR activities DG.RSTL.050.008, ICT.P10.009.001 and the Austrian Science Fund (FWF) grant no. P20134-N13.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 34–41, 2011.

or when data have multidimensional features, such as colors in the RGB model or the coordinate of a point in the 3-dimensional space. These considerations drew the attention to the study of a multidimensional setting [2,6,13]. The term multidimensional, or equivalently k-dimensional, is related to considering measuring functions taking values in $\mathbb{R}^k$, that is, $\boldsymbol{\varphi} : X \to \mathbb{R}^k$, and the subsequent extension of shape descriptors to this case.

As a first solution, in [2] the authors studied the concept of *k-dimensional size functions* and defined the *k-dimensional matching distance* to compare k-dimensional size functions. Unfortunately, they did not explain how to use the latter in practice, namely, how to approximate it so as to obtain a good compromise between computational costs and quality of results. In [5], an algorithm was presented to approximate the k-dimensional matching distance when $k = 2$, up to an arbitrary error threshold.

The contribution of the paper. Our goal is to validate the framework proposed in [5] to deal with 2-dimensional Size Theory. In this sense, the main contributions of the present paper are the following ones:

– We show the robustness of our framework with respect to non-rigid shape deformations. To this aim, we perform an experiment on the database used in the Non rigid world Benchmark [1], which is suited for non-rigid shape retrieval and comparison;

– We show the capability of our framework to deal with other classes of shape deformations, such as non-metric-preserving transformations. To achieve this task we build, starting from the previous database, a new one of 228 models and exploit the modularity of 2-dimensional size functions: We show that, simply by changing the 2-dimensional measuring function, 2-dimensional size functions gain different invariance properties, better suited for this new problem;

– We show how the cited framework can improve the ability of 2-dimensional size functions in shape discrimination, allowing us to tune computational costs and accuracy of results.

The paper is organized as follows. We first overview the main definitions and properties about Size Theory, with particular reference to the 1-dimensional (Section 2) and the 2-dimensional setting (Section 3). Our experiments are shown in Section 4. Some discussions in Section 5 conclude the paper.

2 1-Dimensional Size Functions

Size functions are shape descriptors that code the topological evolution of the sublevel sets of a space X, according to the increasing values of a real function $\varphi : X \to \mathbb{R}$ defined on it, and called 1-*dimensional measuring function.* Indeed, size functions count the number of connected components which remain disconnected passing from a lower level set of X, $X_u = \{P \in X : \varphi(P) \leq u\}$, to another. Since the sequence of lower level sets is driven by the real function φ, size functions encode the geometrical properties of X captured by φ in the topological evolution of X_u. More formally,

Definition 1. *Given a pair* (X, φ) *with* X *a non-empty, compact and locally connected Hausdorff space and* φ *a continuous function, and denoting* $\Delta^+ = \{(u,v) \in \mathbb{R} \times \mathbb{R} : u < v\}$*, the* size function *of* (X, φ) *is* $\ell_{(X,\varphi)} : \Delta^+ \to \mathbb{N}$*, with* $\ell_{(X,\varphi)}(u,v)$ *equal to the number of connected components of the sublevel set* $X_v = \{P \in X : \varphi(P) \le v\}$*, containing at least one point of the sublevel set* X_u*.*

Figure 1(b) shows a simple example of a 1-dimensional size function (1SF), i.e. when the measuring function is real-valued. On the left (Figure 1(a)) the considered pair (X, φ) can be found, where X is the curve drawn by a solid line, and φ is the ordinate function. For example, let us compute the value of $\ell_{(X,\varphi)}$ at the point (c, d). By applying Definition 1, it is sufficient to count how many of the three connected components in the sublevel X_d contain at least one point of X_c. It can be easily checked that $\ell_{(X,\varphi)}(c, d) = 2$.

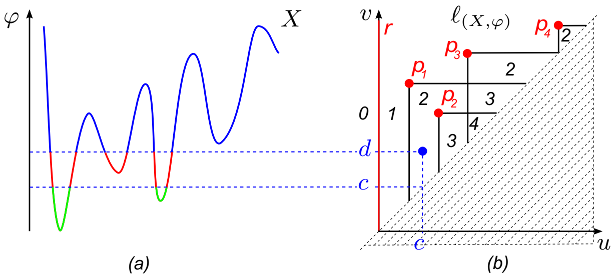


Fig. 1. (a) The space X and a filtering function φ. (b) The associated size function $\ell_{(X,\varphi)}$, represented by the formal series $r + p_1 + p_2 + p_3 + p_4$.

As Figure 1 shows, 1SFs have a typical structure: They are linear combinations (with natural numbers as coefficients) of characteristic functions of triangular regions. That implies that each 1SF can be described by a *formal series*, i.e. a (formal) linear combination of *cornerpoints* (e.g., the points $p_1, \dots, p_4$ in Figure 1(b)) and cornerlines (e.g., the line r in Figure 1(b)). Due to this kind of representation, the original issue of comparing shapes can be turned into a simpler algebraic problem: Each distance between formal series naturally produces a distance between 1SFs, among which the so called *matching distance* [3,9] (also known as *bottleneck distance*, see [11]).

3 2-Dimensional Size Functions

2-dimensional size functions (2SFs) generalize the 1-dimensional situation to the case of measuring functions taking values in $\mathbb{R}^2$, i.e. $\boldsymbol{\varphi} = (\varphi_1, \varphi_2)$.

Extending Definition 1 to the 2-dimensional case is straightforward. For every $\boldsymbol{u} = (u_1, u_2), \boldsymbol{v} = (v_1, v_2) \in \mathbb{R}^2$, we say that $\boldsymbol{u} \preceq \boldsymbol{v}$ (resp. $\boldsymbol{u} \prec \boldsymbol{v}$) if and only if $u_i \le v_i$ (resp. $u_i < v_i$) for $i = 1, 2$. Then we have:

Definition 2. *Let be $\Delta^+ = \{(\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^2 \times \mathbb{R}^2 : \boldsymbol{u} \prec \boldsymbol{v}\}$. The* 2-dimensional size function *of $(X, \boldsymbol{\varphi})$ is $\ell_{(X,\boldsymbol{\varphi})} : \Delta^+ \to \mathbb{N}$, with $\ell_{(X,\varphi)}(\boldsymbol{u}, \boldsymbol{v})$ equal to the number of connected components of the sublevel set $X_{\boldsymbol{v}} = \{P \in X : \boldsymbol{\varphi}(P) \preceq \boldsymbol{v}\}$, containing at least one point of the sublevel set $X_{\boldsymbol{u}}$.*

Reduction to the 1-dimensional case. Since every 1SF may be seen as a linear combination of triangles (formal series of function characteristics), the comparison of two 1SFs is simple and computationally efficient [9]. Unfortunately, the same representation seems not to hold for 2SFs (and kSFs in general).

The solution in [2] was to decompose the domain $\Delta^+ \subset \mathbb{R}^2 \times \mathbb{R}^2$ into a family of half-planes, and prove that the restriction of a 2SF to each of these half-planes is a particular 1SF. As a consequence, a 2SF can be represented by a collection of 1SFs (one for each considered half-plane), and the 1-dimensional matching distance can be applied to every half-plane of the family. This leads to the following definition of a distance between 2SFs:

$$(1) \qquad D_{match}\left(\ell_{(X,\boldsymbol{\varphi})}, \ell_{(Y,\boldsymbol{\psi})}\right) = \sup_{h \in H} d_h\left(\ell_{(X,\boldsymbol{\varphi})}, \ell_{(X,\boldsymbol{\psi})}\right),$$

with H the set of half-planes in $\mathbb{R}^2 \times \mathbb{R}^2$, and $d_h\left(\ell_{(X,\boldsymbol{\varphi})}, \ell_{(Y,\boldsymbol{\psi})}\right)$ the (1-dimensional) matching distance (multiplied by a suitable scaling factor) between the 1SFs on half-plane h [2].

Algorithm for approximating $\boldsymbol{D_{match}}$. Equation (1) implies that, in general, a direct computation of $D_{match}\left(\ell_{(X,\boldsymbol{\varphi})}, \ell_{(Y,\boldsymbol{\psi})}\right)$ is not possible, since we should calculate the value $d_h\left(\ell_{(X,\boldsymbol{\varphi})}, \ell_{(X,\boldsymbol{\psi})}\right)$ for an infinite number of half-planes h. On the other hand, if we choose a non-empty and finite subset $S \subseteq H$, and substitute $\sup_{h \in H}$ with $\max_{h \in S}$ in Equation (1), we get a computable distance, say $\mathscr{D}_{match}\left(\ell_{(X,\boldsymbol{\varphi})}, \ell_{(Y,\boldsymbol{\psi})}\right)$, that can be effectively used in concrete applications. Thinking of $\mathscr{D}_{match}\left(\ell_{(X,\boldsymbol{\varphi})}, \ell_{(Y,\boldsymbol{\psi})}\right)$ as an approximation of $D_{match}\left(\ell_{(X,\boldsymbol{\varphi})}, \ell_{(Y,\boldsymbol{\psi})}\right)$, we can argue that the larger the set $S \subseteq H$, the smaller the difference between the two values. On the other hand, the smaller the set S, the faster the computation of $\mathscr{D}_{match}$. In this perspective, in [5] the authors presented an algorithm to find a set S that is a compromise between these two situations. Taking as input an arbitrary real value $\varepsilon > 0$ which plays the role of an error threshold, the proposed procedure looks for a suitable set S, giving as output an approximation $\mathscr{D}_{match}\left(\ell_{(X,\boldsymbol{\varphi})}, \ell_{(Y,\boldsymbol{\psi})}\right)$ of $D_{match}\left(\ell_{(X,\boldsymbol{\varphi})}, \ell_{(Y,\boldsymbol{\psi})}\right)$ satisfying the relation

$$(2) \qquad 0 \le D_{match}\left(\ell_{(X,\boldsymbol{\varphi})}, \ell_{(Y,\boldsymbol{\psi})}\right) - \mathscr{D}_{match}\left(\ell_{(X,\boldsymbol{\varphi})}, \ell_{(Y,\boldsymbol{\psi})}\right) \le \varepsilon.$$

The algorithm follows an iterative, multi-scale approach based on some theoretical results enabling to bound the changing of $d_h\left(\ell_{(X,\boldsymbol{\varphi})}, \ell_{(X,\boldsymbol{\psi})}\right)$ according to the choice of the half-planes in the considered collection [5, Lemma 3.1 and Thm. 3.4]. As a by-product, these results can be used to define a cancellation strategy and get rid of a large number of half-planes throughout the requested iterations, thus speeding up the procedure at any rate. This is the algorithm we use for the experiments in this paper.

4 Experimental Results

Our goal is to validate the computational framework roughly described in the previous section and proposed in [5] to deal with 2-dimensional Size Theory. Through some experiments on 3D-models represented by triangle meshes, we will prove that the cited framework is robust with respect to noise, as well as to non-rigid and non-metric-preserving shape transformations. To this aim we will make use of the modularity property of 2-dimensional size functions: It will be shown that, in order to deal with invariance to different classes of transformations, it is sufficient to change the considered 2-dimensional measuring functions, without changing anything else in the framework. Our experiments will finally make clear that the proposed procedure can actually improve the ability of 2SFs in shape discrimination, allowing us to tune computational costs and accuracy of results.

Computational aspects. From the computational point of view, the reduction of 2SFs to the 1-dimensional case allows us to use the existing framework for computing 1SFs. In this discrete (1-dimensional) setting, the counterpart of a pair (X, φ) is given by a *size graph* (G, φ), where $G = (V(G), E(G))$ is a finite graph, with $V(G)$ and $E(G)$ the set of vertices and edges respectively, and $\varphi : V(G) \to \mathbb{R}$ is a measuring function defined on the nodes of the graph [10].

In our experiments, the size graph is made of the vertices and the edges of the triangle mesh. Once the size graph has been built, computing the restriction of a 2-dimensional size function on a single half-plane takes $O(n \log n + \alpha(2m + n, n))$, where n and m are the number of vertices and edges in the size graph, respectively, and α is the inverse of the Ackermann function [10].

As stressed before, the algorithm we implement in our experiments is based on an iterative approach, the number of iteration proportional to the accuracy in results we want to achieve. This accuracy can be fixed by choosing the threshold error ε, which gives us the maximum distance between the output, i.e. the value $\mathscr{D}_{match}\left(\ell_{(X,\boldsymbol{\varphi})}, \ell_{(Y,\boldsymbol{\psi})}\right)$ and the actual 2-dimensional matching distance $D_{match}\left(\ell_{(X,\boldsymbol{\varphi})}, \ell_{(Y,\boldsymbol{\psi})}\right)$, cf equation (2).

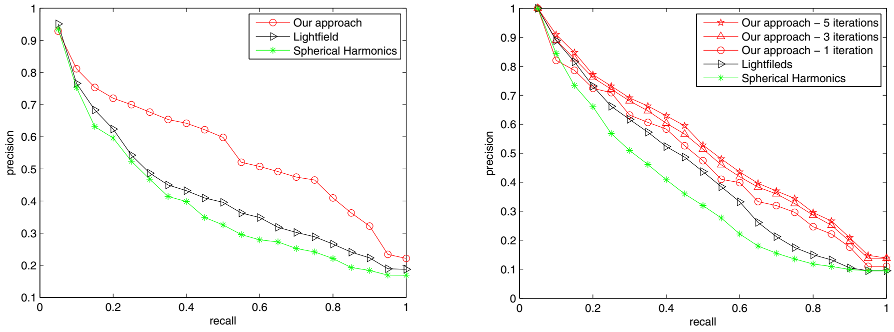
Table 1 shows some statistics concerning the average time required to compute and compare the 2-dimensional size functions associated to different 3D-models, in comparison with an accuracy evaluation of results. In particular, the value of the threshold error ε is expressed in accuracy percentage points, i.e. by computing the ratio $\mathscr{D}_{match}\left(\ell_{(X,\boldsymbol{\varphi})}, \ell_{(Y,\boldsymbol{\psi})}\right) / (\mathscr{D}_{match}\left(\ell_{(X,\boldsymbol{\varphi})}, \ell_{(Y,\boldsymbol{\psi})}\right) + \varepsilon)$ for every possible comparison, and then taking the average. These results have been obtained on a 2.8GH Core i5, RAM 8GB.

Table 1. Statistics on average time for computing and comparing 2 different 2Sfs

Iterations (number)	1	2	3	4	5	6
Accuracy (%)	22%	36%	53%	69%	82%	90%
average time (seconds)	0.34	1.32	2.27	7.09	22.13	81.71

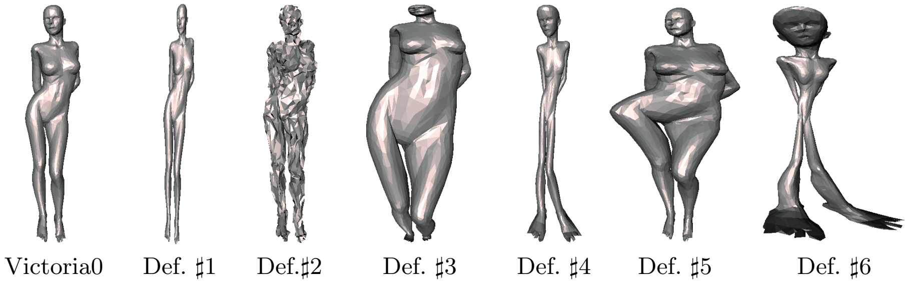
Non-rigid shape similarity. To analyze the potential of the proposed method for comparing and retrieving shapes, we performed some tests on the database of 148 triangle meshes used in the Non-rigid world Benchmark [1]. Since this database is suited for non-rigid shape retrieval and comparison purposes, it would be desirable to have shape descriptors that are robust with respect to non-rigid shape deformations. To this aim, we can exploit a fundamental aspect of 2-dimensional size functions: They inherit their invariance properties (w.r.t. to groups of transformations) directly from the associated measuring functions. Therefore, in this experiment we selected the following 2-dimensional measuring functions: The first component is chosen to be the heat kernel signature [17], computed using the first 10 eigenfunctions of the Laplace-Beltrami operator and a fixed time $t = 1000$, and the second one the integral geodesic distance [14]. Indeed, it is well known that this two functions are robust with respect to non-rigid shape changing. The obtained results (after 4 iterations of the algorithm) can be seen in the precision-recall graph in Table 2(a). The considered query set coincides with the all database. We also compared our framework with two methods representing a state-of-the-art shape retrieval techniques: the Spherical harmonics (SH) method [15] and the Light Field Distribution (LFD) method [8].

Table 2. On the left (Table2(a)), the average precision-recall graphs concerning non-rigid shape similarity. On the right (Table2(b)), the average precision/recall graphs of our framework at different levels of accuracy, concerning robustness experiments.



Robustness to noise and non-metric-preserving deformations. To test the robustness of the proposed framework we built a new database of 228 3D-surface mesh models. We started by considering one model for each of the 12 class of the dataset used in the Non rigid world Benchmark (cat0, david0, dog0,..., victoria0, wolf0). For each model, six non-rigid transformations were applied, at three different strength levels. An example of the transformations and their strength levels is given in Table 3. What we get at the end is a database containing 12 classes, each one consisting of a null model together with its 18 modified versions. We created a new database mainly because we want to emphasize the robustness of our method not only with respect to non-rigid transformations, as

Table 3. The null model "Victoria0" and the 3^{rd} strength level for each deformation



shown in the previous experiment, but also with respect to noise (e.g Deformation ♯2) and other deformations not subject to metric constraints: Actually, we consider transformations which do not preserve the metric properties of shapes (e.g. the Riemannian metric).

The previous considerations imply that the descriptors used to compare non-rigid shapes could not be enough for this new task. But fortunately we can rely on the modularity of 2-dimensional size functions: To obtain different invariance properties, we simply have to change the 2-dimensional measuring function. For each triangle mesh of vertices $\{P_1, \ldots, P_n\}$, we define a new 2-dimensional measuring function as follows. We compute the center of mass B, and normalize the model so that it is contained in a unit sphere. We also define a vector

$$\boldsymbol{w} = \frac{\sum_{i=1}^{n}(P_i - B)\|P_i - B\|}{\sum_{i=1}^{n}(P_i - B)\|P_i - B\|^2}.$$

The 2-dimensional measuring function $\boldsymbol{\varphi} = (\varphi_1, \varphi_2)$ is then chosen such that φ_1 is the distance from the line parallel to $\boldsymbol{w}$ and passing through B, and φ_2 is the distance from the plane orthogonal to $\boldsymbol{w}$ and passing through B. The values of φ_1, φ_2 are finally normalized so that they range in the interval $[0, 1]$. Note that both φ_1 and φ_2 are invariant with respect to translation and rotation, while the invariance to scale comes from the a priori normalization of the models.

Table 2(b) shows the average precision/recall graphs of our framework at different levels of accuracy, i.e. after a different number of iterations of the proposed procedure, and in comparison with the (SH) and the (LFD) methods. The query set coincides with the all database. We emphasize that our results improve proportionally to the number of iterations of our algorithm.

5 Conclusions

In this paper we validated a new framework presented in [5] to compute an approximation of the matching distance between 2-dimensional size functions. More precisely, we proposed experiments proving that the cited framework is modular and robust with respect to noisy, non-rigid and non-metric-preserving shape transformations. Our results show how the cited framework can be used to improve the ability of 2-dimensional size functions in comparing shapes, allowing us to find a good compromise between computational costs and goodness of results. For the next future, it could be interesting to study how to extend the discussed procedure to higher dimensional settings of Size Theory.

References

1. http://tosca.cs.technion.ac.il/
2. Biasotti, S., Cerri, A., Frosini, P., Giorgi, D., Landi, C.: Multidimensional size functions for shape comparison. J. Math. Imaging Vision 32(2), 161–179 (2008)
3. Biasotti, S., De Floriani, L., Falcidieno, B., Frosini, P., Giorgi, D., Landi, C., Papaleo, L., Spagnuolo, M.: Describing shapes by geometrical-topological properties of real functions. ACM Comput. Surv. 40(4), 1–87 (2008)
4. Biasotti, S., Giorgi, D., Spagnuolo, M., Falcidieno, B.: Size functions for comparing 3d models. Pattern Recogn. 41(9), 2855–2873 (2008)
5. Biasotti, S., Cerri, A., Frosini, P., Giorgi, D.: A new algorithm for computing the 2-dimensional matching distance between size functions. Tech. Rep. no. 2821,Universitá di Bologna (2010), http://amsacta.cib.unibo.it/2821/
6. Carlsson, G., Zomorodian, A.: The theory of multidimensional persistence. In: SCG 2007, pp. 184–193 (2007)
7. Cerri, A., Ferri, M., Giorgi, D.: Retrieval of trademark images by means of size functions. Graph. Models 68(5), 451–471 (2006)
8. Chen, D.Y., Tian, X.P., Shen, Y.T., Ouhyoung, M.: On visual similarity based 3d model retrieval. Comput. Graph. Forum 22(3), 223–232 (2003)
9. d'Amico, M., Frosini, P., Landi, C.: Natural pseudo-distance and optimal matching between reduced size functions. Acta. Appl. Math. 109, 527–554 (2010)
10. d'Amico, M.: A New Optimal Algorithm for Computing Size Functions of Shapes. In: CVPRIP Algorithms III, Atlantic City, pp. 107–110 (2000)
11. Edelsbrunner, H., Harer, J.: Computational Topology: An Introduction. American Mathematical Society, Providence (2009)
12. Frosini, P., Landi, C.: Size functions and formal series. Appl. Algebra Engrg. Comm. Comput. 12(4), 327–349 (2001)
13. Frosini, P., Mulazzani, M.: Size homotopy groups for computation of natural size distances. Bulletin of the Belgian Mathematical Society 6(3), 455–464 (1999)
14. Hilaga, M., Shinagawa, Y., Kohmura, T., Kunii, T.L.: Topology matching for fully automatic similarity estimation of 3d shapes. In: SIGGRAPH 2001, pp. 203–212 (2001)
15. Kazhdan, M., Funkhouser, T., Rusinkiewicz, S.: Rotation invariant spherical harmonic representation of 3d shape descriptors. In: Proc. SGP 2003, pp. 156–164 (2003)
16. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. PAMI 22(12) (2000)
17. Sun, J., Ovsjanikov, M., Guibas, L.: A concise and provably informative multi-scale signature based on heat diffusion. In: Proc. SGP 2009, pp. 1383–1392 (2009)
18. Tangelder, J., Veltkamp, R.: A survey of content-based 3D shape retrieval methods. Multimedia Tools and Applications 39(3), 441–471 (2008)

A Homological–Based Description of Subdivided nD Objects

Helena Molina-Abril[1,2] and Pedro Real[1,⋆]

[1] Computational Topology and Applied Mathematics Group, University of Seville
habril@us.es, real@us.es
[2] Pattern Recognition and Image Processing Group, Vienna University of Technology

Abstract. We present here a topo–geometrical description of a subdivided nD object called homological spanning forest representation. This representation is a convenient tool in order to completely control not only geometrical, but also advanced topological information of a given object. By codifying the underlying algebraic topological machinery in terms of coordinate–based graphs, we progress in the task to "combinatorialize" homological information at two levels: local and global. Therefore, our method presents several advantages with respect to the existing Algebraic topological models, and techniques based in Discrete Morse Theory. A construction algorithm has been implemented, and some examples are shown.

1 Introduction

One way to guarantee a consistent description of an object is to base such description on topological principles. A topological representation of an object defines a finite topological space made up of regions, arcs, and points which encode a particular partitioning. Several structures have been proposed to encode such a partitioning, including cellular complexes [1,2], combinatorial maps [3], graphs [4], etc.

We deal here with the problem of finding an efficient and robust geometrical and topological representation of a subdivided nD object given in terms of a cell complex and exploiting the notion of homology or, more precisely, using chain homotopy equivalences connecting the object with its homology groups (see [5] for more details).

In principle, homology is a purely algebraic notion related to the degree of connectivity at the level of formal sum of cells (connected components, holes or tunnels, cavities,...) and most of the models based on these ideas are algebraic–topological models (AT–model [6], AM–model [7],....). Nevertheless, it is possible to "combinatorialize" these models (eliminate its algebraic part) by using a graph representation of the algebraic operators (boundary operator, coboundary

[⋆] This work has been partially supported by the Andalusian Research Group CATAM FQM-296, the Spanish MICINN Research Project MTM2009-12716 and the Austrian Science Fund under grant P20134-N13.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 42–50, 2011.

operator,...) and simplify its connectivity information by using hierarchical tree–like structures. Forman [8,9] used this idea in order to develop Discrete Morse Theory (DMT, for short), which has become a powerful tool in its applications to computational topology, computer graphics, image processing and geometric modeling.

With the philosophy of representing the object in terms of a finite number of topologically inessential "threads" as a goal, we translate DMT to a suitable algebraic homological setting, that of integral–chain complexes. We can progress in this way in the task to "combinatorialize" homological information at two level: local (DMT) and global (in terms of coordinated–based hierarchical forests based on chain homotopies [10]). The algebraic nature of the global approach can be combinatorialized if we place these chain homotopies and critical cells (homology generators) in a graph–based ambiance. Based on that, we develop here a non–unique combinatorial topo–geometric representation of a nD subdivided object, called homological spanning forest (or HSF, for short). We design and implement an algorithm for computing this model for a nD object embedded in $\mathbb{R}^n$ and we do some experiments in the three dimensional case.

2 Preliminaries

In this section, we first establish a notion of (combinatorial) *cell complex* in a finite–dimensional Euclidean space with the cell boundary information described in algebraic terms.

A *cell complex* $K = \{K_i\}_{i=0}^{\ell}$ embedded in $\mathbb{E}^{\ell}$ is a finite collection of cells $\{\sigma_i^{(r)i=1,\ldots n} \in K_r\}$ of different dimensions $0 \geq r \geq \ell$ such that (see [11] for a formal definition of cell):

(i) $|K| = \bigcup_{i=1}^{n} \sigma_i = |K_0| \cup |K_1| \cup \ldots \cup |K_\ell|$. The set K_r consists of all the r–cells of K, for $0 \leq r \leq \ell$. It is possible that $K_i = \emptyset$ for some $0 < i \leq \ell$.

(ii) $\sigma_i \cap \sigma_j = \emptyset \ (i \neq j)$;

(iii) If $dim(\sigma_i) = p$ (with $0 \geq p \geq \ell$), then $\partial\sigma_i \subset \bigcup_{i=1}^{p-1} K_i$,

The *p–skeleton* $K^{(p)}$ for K is the set of all k–cells with $0 \leq k \leq p$. The dimension of the cell complex is the smallest non–negative natural number r such that the condition $K^{(r)} = K^{(r+1)}$ is satisfied. If all the cells of K are convex sets of $\mathbb{E}^{\ell}$, then K is called *convex cell complex*. Simplicial, Cubical and some polyhedral complexes are special cases of convex cell complexes.

Roughly speaking, the idea of homology is to analyze the degree of connectivity of cell complexes using formal sums of cells. A *differential operator* for a cell complex K with coefficients in Λ is a linear map $d : \Lambda[K] \to \Lambda[K]$, such that the image of a p–cell σ is a linear combination of some $(p-1)$–cells of the boundary $\partial(\sigma)$ and $d \circ d = 0$. Taking into account that our cell complex K is embedded in $\mathbb{E}^{\ell}$, its geometric realization $|K|$ is a regular triangulable cell complex and there can be always defined a differential operator ∂, called *boundary operator*, with coefficients in the field Λ.

The chain complex canonically associated to the cell complex K is the graded differential vector space $(C_*(K), \partial)$, where $C_p(K) = \Lambda[K_p]$, for all $p = 0, 1, \dots r$, and $\partial : C_*(K) \to C_{*-1}(K)$ is the previous boundary operator for the cell complex K. For instance, to find a boundary operator ∂ for a simplicial complex is straightforward, but it is not, in general, an easy task for others cell complexes. The following is one of the fundamental results in the theory of CW–complexes.

Theorem 1 (See [8]). *Let K a finite cell complex. There are algebraic boundary maps $\partial_p : C_p(K, \Lambda) \to C_{p-1}(K, \Lambda)$, for each p, so that $\partial_{p-1} \circ \partial_p = 0$ and such that the resulting differential complex $\{C_p(K, \Lambda), \partial_p\}_{p=0}^r$ calculates the homology of K. That is, if we define $H_p(C, \partial) = \text{Ker } (\partial_p)/\partial_{p+1}(C)$. In other words, $H_p(C, \partial) \cong H_p(|K|, \Lambda)$.*

From now on, a finite cell complex K is denoted by (C, ∂), where $\partial : C_*(K) \to C_{*-1}(K)$ is the boundary operator for $C_*(K)$ with coefficients in the finite field $\mathbb{Z}/\mathbb{Z}2 = \{0, 1\}$.

3 Integral–Chain Complexes

In [12], we recover the algebraic machinery underlying in Discrete Morse Theory, establishing a new framework for dealing with special chain complexes, that is the integral–chain complexes associated to finite cell complexes. In this section, we recall the main notions and results of this homological algebra work in order to understand our approach.

Definition 1. *[12] An* integral chain complex *(C, ∂, ϕ) is a graded vector space $C = \{C_p\}_{p=0}^n$ endowed with two linear maps: a differential operator $\partial : C_* \to C_{*+1}$, and an integral operator (also called* algebraic gradient vector field *[8] or* chain homotopy operator *[5]) $\phi : C_* \to C_{*+1}$, satisfying the global nil potency properties $\partial \circ \partial = 0$ and $\phi \circ \phi = 0$. An integral chain complex (C, ∂, ϕ) is ∂–*pure *(resp. ϕ–*pure*) if the condition $\partial = \partial \circ \phi \circ \partial$, called homology condition (resp. the condition $\phi = \phi \circ \partial \circ \phi$, called* strong deformation retract *condition) is satisfied.*

Examples of the application of integral operators is shown in Figure 1.

The computation of the homology of a chain complex (C, ∂) can be specified in terms of finding an integral operator $\phi : C_* \to C_{*+1}$, satisfying the Strong Deformation Retract (SDR for short) and homology conditions with regards to the differential operator ∂ ([13]).In spite of its simplicity, the following result is essential.

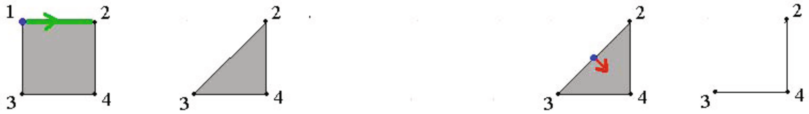


Fig. 1. A cell complex and the resulting cell complex after applying the integral operator $\phi(\langle 1 \rangle) = \langle 1, 2 \rangle$ (on the left) and a cell complex and the result after applying $\phi(\langle 1, 2 \rangle) = \langle 1, 2, 3 \rangle$ (on the right)

Lemma 1. *[Integral–Chain Lemma] An integral chain complex (C, ∂, ϕ) is integral–chain equivalent to its harmonic complex $\pi(C, \partial, \phi)$, meaning that its homological information can be extracted from that of the harmonic (see [12]). This last harmonic complex $\pi(C, \partial, \phi)$ is of the form $(\pi(C), \partial_\pi, \phi_\pi)$ where $\pi(x) = (id + \phi \circ \partial + \partial \circ \phi)(x)$, $\partial_\pi(\pi(x)) = (\partial - \partial \circ \phi \circ \partial)(x)$ and $\phi_\pi(\pi(x)) = (\phi - \phi \circ \partial \circ \phi)(x)$. $\forall x \in C$.*

We now define an algebraic constructor of a new integral–chain complex:

Definition 2. *Given a ∂–pure (resp. two ϕ–pure) integral–chain complexes (C, ∂, ϕ) and a differential operator ∂' satisfying the homology condition (resp. an integral operator ϕ' satisfying the strong deformation retract condition) for $\pi_{(\partial,\phi)}(C)$, a new ∂–pure (resp. ϕ–pure) integral chain complex $(C, \partial + \partial' \circ \pi_{(\partial,\phi)}, \phi)$ (resp. $(C, \partial, \phi + \phi' \circ \pi_{(\partial,\phi)})$) can be constructed. This new integral chain complex is called* composition *of (C, ∂, ϕ) by ϕ'.*

From now on, all the integral chain complexes we consider in the paper will be ϕ–pure integral–chain complexes.

4 Homological Spanning Forest Representation

Discrete Morse Theory (DMT, for short) gives a positive answer to the problem of finding combinatorial chain homotopy operators ϕ for chain complexes $(C_*(K), \partial)$ of finite cell complexes, such that the integral homology of $(C_*(K), \partial, \phi)$ is a "good" approximation (measured in terms of critical cells, determined by the Betti numbers) to its differential homology. These combinatorial integral operators are seen in DMT as cell pairings (cell collapses) (see Figure 2). In [14], given a 2–manifold, an heuristic for computing optimal Morse pairings is developed. A pairing is considered optimal when the discrete gradient vector field has few critical cells (cells that are not paired). This heuristic computes optimal gradient vector fields in terms of hyper-forests. However, for general cell complexes this problem has not been solved yet.

Now, we progress in DMT with some slightly modifications with regards the classical theory, and without using, in principle, discrete Morse functions. In this way, we are able to obtain an optimal pairing for any finite cell complex without restriction.

Definition 3. *A combinatorial integral operator $\mathcal{V}$ defined on a cell complex K is a collection of disjoint pairs of (non–necessarily incident) cells $\{\alpha^{(p)} \prec \beta^{(p+1)}\}$. If the pairs are constituted by incident cells then, $\mathcal{V}$ is called combinatorial vector field ([8]). A cell α is a critical cell of $\mathcal{V}$ if it is not paired with any other cell in $\mathcal{V}$.*

In the sequel, we prefer to describe some important notions in terms of the barycentric subdivision of a cell complex rather than using its Hasse diagram.

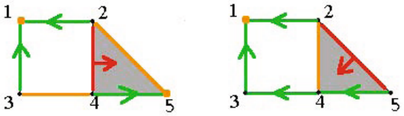


Fig. 2. A cell pairing on the left (⟨1⟩,⟨5⟩, ⟨3, 4⟩ and ⟨2, 5⟩ are critical), and an optimal one on the right (⟨1⟩ and ⟨2, 4⟩ are critical). The pairing is represented with an arrow from the cell of lower dimension to its paired cell of higher dimension.

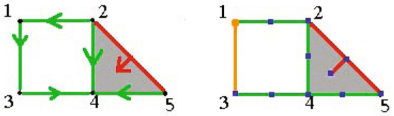


Fig. 3. A combinatorial vector field (on the left). On the right a gradient set of forest where cells ⟨1⟩ and ⟨1, 3⟩ do not belong to the forest, ⟨2, 5⟩ and ⟨2, 4, 5⟩ belong to F_1 and the rest of cells belong to F_0.

Definition 4. *Let (K, ∂) be a finite convex cell complex of dimension m embedded in $\mathbb{R}^n$ and let $(BCS(K), \partial^{bcs})$ be the simplicial barycentric subdivision of K [15]. Let us consider a hierarchy of simplicial forests $\mathcal{F} = \{F_0, F_1, \ldots, F_{m-1}\}$ all contained in the 1–skeleton of $BCS(K)$, such that the nodes of F_p are vertex cells of dimension p and $p + 1$ of K and its leaves are cells of dimension p, for all $0 \leq p \leq m$. Such set of forests is called gradient set of forests for the cell complex (K, ∂).*

An example of a combinatorial vector field and a gradient set of forest is shown in Figure 3.

A gradient set of forests $\mathcal{F}$ for (K, ∂) can be expressed in combinatorial terms by means of a combinatorial vector field $\mathcal{V}_{\mathcal{F}}$, called *$\mathcal{F}$–gradient vector field.* In fact, $\mathcal{V}_{\mathcal{F}}$ is defined by choosing a root in each tree T of $\mathcal{F}$ and pairing incident cells (of different dimension) of T excepting the root. In this way, the roots of the trees become critical cells of $\mathcal{V}_{\mathcal{F}}$ as well as the rest of cells in K which do not appear in $\mathcal{F}$.

A ϕ–pure integral–chain complex $(C(K), \partial, \tilde{\mathcal{V}_{\mathcal{F}}})$ can be derived from $(C(K), \partial, \mathcal{V}_{\mathcal{F}})$. If $\sigma^{(p)}$ is a node of $\mathcal{F}$ which is not a root, $\tilde{\mathcal{V}_{\mathcal{F}}}(\sigma^{(p)})$ is defined as the sum of the $(p+1)$–cells of K existing in the unique path within the forest F_p, joining σ and the corresponding root. In other case, its value is zero. This *ϕ–pure $\mathcal{F}$–gradient integral operator* $\tilde{\mathcal{V}_{\mathcal{F}}}$ satisfies the SDR–condition $\mathcal{V}_{\mathcal{F}} \partial \mathcal{V}_{\mathcal{F}} = \mathcal{V}_{\mathcal{F}}$.

The main algorithm of this paper is designed using as main piece the following proposition which is already proved in [12].

Proposition 1. *[12] Let (K, ∂) be a finite convex cell complex and let $\mathcal{F}$ a gradient set of forest for K. Let $(C(K), \partial, \phi)$ be the integral–chain complex, being ϕ the pure gradient integral operator derived from $\mathcal{F}$. Then, the harmonic complex* *of* $(C(K), \partial, \tilde{\mathcal{V}_{\mathcal{F}}})$ *is isomorphic to* $(\text{ Ker } \mathcal{V}_{\mathcal{F}} \setminus \mathcal{V}_{\mathcal{F}}(C(K)), \partial_\pi, 0)$*. This last integral–chain complex, called Morse cell complex* $\mathcal{M}(C(K), \mathcal{F})$ *is constituted by finite linear combinations of the different critical cells of* $\mathcal{V}_{\mathcal{F}}$ *and* ∂_π *can be seen as the boundary operator of the corresponding cell complex determined by the critical cells. Given a critical p–cell* $\sigma^{(p)}$*, then* $\partial_\pi(\sigma^{(p)}) = (\partial - \partial \circ \phi \circ \partial)(\sigma^p)|_{criticalcells}$*, that is, the linear combination of the critical* $(p-1)$*–cells appearing in* $(\partial - \partial \circ \phi \circ \partial)(\sigma^{(p)})$*.*

Let us note that we can repeatedly apply Prop. 1 to the successive Morse complexes, previously describing a corresponding gradient forest for each of them. We can express this task in the following way:

$$(C(K), \partial) \Rightarrow \mathcal{M}(C(K), \mathcal{F}^0) \Rightarrow \mathcal{M}(\mathcal{M}(C(K), \mathcal{F}^0), \mathcal{F}^1) \Rightarrow$$
$$M(\mathcal{M}(\mathcal{M}(C(K), \mathcal{F}^0), \mathcal{F}^1), \mathcal{F}^2) \Rightarrow \ldots \Rightarrow H_*(C(K), 0),$$

where $(C, \partial) \Rightarrow (C', \partial')$ means that there is a chain homotopy equivalence [15] between the chain complexes (C, ∂) and (C', ∂').

On the other hand, Prop. 2 provides us the integral–chain complex $(C(K), \partial, h)$ which is composition of $(C(K), \partial, 0)$ by the successive gradient forests. If we obtain in this way a Morse complex, with all the possible gradient forest being trivial (constituted of only one node), the process stops and h is the "key" operator for getting the homology groups and corresponding homology generators of K as well as a topological interpretation of K in terms of trees in the 1–skeleton of the barycentric subdivision of K. This geo–topological (coordinate–based) representation is called *Homological Spanning Forest* (or HSF, for short) representation for K (see Figure 4).

Theorem 2. *In the previous conditions, the integral operator* $h : C_*(K) \rightarrow C_{*+1}(K)$ *specifies a set of forest* $\mathcal{G} = \{G_0, G_1, \ldots, G_m\}$ *in* $BCS(K)$ *called HSF–representation for* K*.*

5 Implementation and Experiments

In [6] the authors present an algorithm to reduce a initial chain complex up to its minimal homological expression. The advantage of this method is that the obtained integral operator encodes the homological information of the initial complex (homology groups, cohomology, homology generators, relations between them, etc.). The complexity in time of this method is $O(n^3)$.

In this section we present a new algorithm, based in Prop. 1, where the resulting HSF representation encodes exactly the same information that the previous mentioned method, and besides the advantages of providing such a representation of the object, by using graph techniques, the time complexity is reduced. The heart of the proposed algorithm runs in linear time, and the question of how many times the loop should be executed, crucially depends on the particular complex.

Given an initial cell complex $C_*(K)$, Algorithm 1 computes its HSF–representation. The algorithm consists of an iterative process, where at each

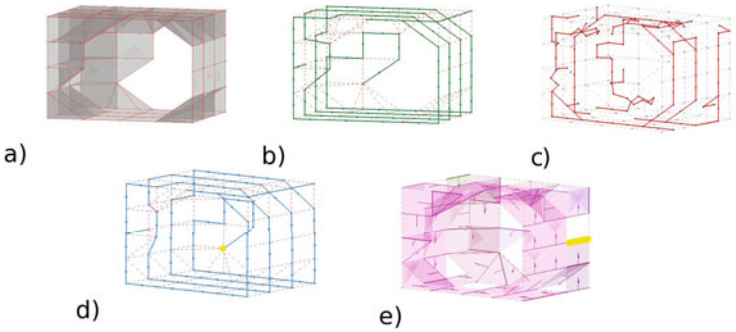


Fig. 4. In Figure 4 a) we can see a cell complex. Part of its homological spanning forest representation (G_0 and G_1) is shown in Figures 4 b) and 4 c). Figures 4 d) and 4 e) represent the optimal combinatorial pairing. The resulting critical cells are colored in yellow in Figures 4 b) and 4 c).

step i, a gradient set of forests $\{F_0 \dots F_p\}$ is computed over $C_*(K)^i$. The function $\mathcal{M}$ returns a Morse complex, constituted by finite linear combinations of the different critical cells in $(C_*(K)^i, \mathcal{F}^i)$. Once the computed $\mathcal{F}^i$ is trivial, the process stops. The guarantee that the minimal number of critic cells is obtained at the end of the algorithm arises in the fact that the algorithm only stops when $\partial = 0$ for every cell in K.

Algorithm 1. HSF$(C_*(K), \partial)$

```
$i = 0$
**while** ! $(Trivial\ (\mathcal{F}^i))$ **do**
    $F_0 = SpanningTree_c\ (C_{0,1}(K)^i)$
    $F_1 = SpanningTree_c\ (C_{1,2}(K)^i \backslash F_0)$
    $\dots$
    $F_p = SpanningTree_c\ (C_{p,p+1}(K)^i \backslash F_{p-1})$
    $\mathcal{F}^i = F_0 \cup F_1 \cup \dots \cup F_p$
    $\mathcal{G}^i = \mathcal{G}^{i-1} \bigcup \mathcal{F}^i$
    $C_*(K)^{i+1} = \mathcal{M}(C_*(K)^i, \mathcal{F}^i)$
    $i = i + 1$
**end while**
**return** $(\mathcal{G}^i)$
```

The union operation $\bigcup$ in Algorithm 1 consists of the integration of the information residing in the forests $\{F_0 \dots F_p\}$ to the global forest $\mathcal{G}$. This itegration is done by using the algebraic composition operation of Prop. 2.

The computation of the gradient set of forests is performed using the Algorithm $SpanningTree_c$. Algorithm $SpanningTree_c$ is a basic spanning tree algorithm, where some extra conditions need to be considered. The basic idea

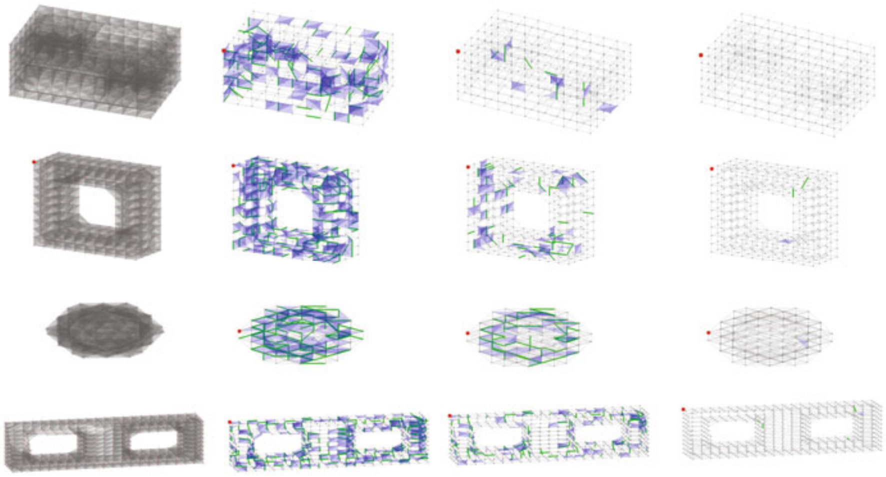


Fig. 5. A Bing's house, Torus, Sphere and Double Torus cell complexes, and their corresponding Morse complexes after some reductions. The final number of critical cells in the final complexes coincides with the Betti Numbers.

of this method, is, to construct valid trees (by joining p–cells and $(p+1)$–cells) that satisfy the global nil potency properties and the SDR condition. Therefore, we must asure that no cycles are created throughout the process.

We have used Depht First Search for the implementation, but any other spanning tree algorithm could be used instead. The implementation is written in C++, and it works either with simplicial or cell complexes. Several experiments have been performed (see Figure 5) using well known examples (Torus, Bing's house, Double Torus, Sphere, etc.). The software has provided valid HSF–representations and the minimum number of critical cells for each example.

6 Conclusions

In this paper we develop a non–unique combinatorial topo–geometric representation of a nD subdivided object, called homological spanning forest (or HSF, for short). This representation is a convenient tool in order to compute not only the minimum number of critical cells but also geometric (local curvature, normals to the boundary, Ricci curvature,....) and advanced topological information (reconstruction of the boundary, homological classification of cycles, relative homology with regards any sub-complex, skeletons, (co)homology operations, . . .). Advantages with respect to existing Algebraic topological models, and DMT–based techniques have been shown. In a near future, we have the intention to deal with the "good" behaviour of the HSF representation for objects embedded in $\mathbb{R}^n$ with regards to combinatorial, geometric and topological changes, simplification, recognition, visualization, etc.

References

1. Kovalevsky, V.: Finite topology as applied to image analysis. Computer Vision, Graphics and Image Processing 46, 141–161 (1989)
2. Klette, R.: Cell complexes through time. Communication and Information Technology Research Technical Report 60 (2000)
3. Edmonds, J.: A combinatorial representation for polyhedral surfaces. Notices Amer. Math. Soc. 7 (1960)
4. Kropatsch, W.G., Haxhimusa, Y., Ion, A.: Multiresolution image segmentations in graph pyramids. In: Applied Graph Theory in Computer Vision and Pattern Recognition, pp. 3–41 (2007)
5. Eilenberg, S., Mac Lane, S.: On the groups $h(\pi, n)$, i, ii, iii. Annals of Math. 58, 60, 60, 55–106,48–139, 513–557 (1953,1954)
6. González-Díaz, R., Real, P.: On the cohomology of 3d digital images. Discrete Appl. Math. 147, 245–263 (2005)
7. González-Díaz, R., Jiménez, M.J., Medrano, B., Real, P.: Chain homotopies for object topological representations. Discrete Appl. Math. 157, 490–499 (2009)
8. Forman, R.: A Discrete Morse Theory for Cell Complexes. In: Yau, S.T. (ed.) Topology and Physics for Raoul Bott. International Press (1995)
9. Forman, R.: Morse theory for cell complexes. Adv. in Math. 134, 90–145 (1998)
10. Molina-Abril, H., Real, P.: Homological computation using spanning trees. In: Bayro-Corrochano, E., Eklundh, J.-O. (eds.) CIARP 2009. LNCS, vol. 5856, pp. 272–278. Springer, Heidelberg (2009)
11. Whitehead, J.: Combinatorial homotopy i. Bull. Amer. Math. Soc. 55, 213–245 (1949)
12. Molina-Abril, H., Real, P.: Homological optimality in discrete morse theory through chain homotopies. Submitted to Pattern Recognition Letters (2011)
13. Gugenheim, V.K.A.M., Lambe, L.A., Stasheff, J.D.: Perturbation theory in differential homological algebra. Illinois J. Math. 33, 357–373 (1989)
14. Lewiner, T., Lopes, H., Tavares, G., Matmídia, L.: Towards optimality in discrete Morse theory. Experimental Mathematics 12 (2003)
15. Munkres, J.: Elements of algebraic topology. Addison Wesley, Reading (1984)

Statistical Shape Model of Legendre Moments with Active Contour Evolution for Shape Detection and Segmentation*

Yan Zhang[1], Bogdan J. Matuszewski[1],
Aymeric Histace[2], and Frédéric Precioso[2,3]

[1] ADSIP Research Centre, University of Central Lancashire
Preston PR1 2HE, UK
{yzhang3,bmatuszewski1}@uclan.ac.uk
[2] ETIS Lab, CNRS/ENSEA/Univ Cergy-Pontoise, 6 av. du Ponceau,
95014 Cergy-Pontoise, France
{frederic.precioso,aymeric.histace}@ensea.fr
[3] LIP6 UMR CNRC 7606, UPMC Sorbonne Universités,
Paris, France
frederic.precioso@lip6.fr

Abstract. This paper describes a novel method for shape detection and image segmentation. The proposed method combines statistical shape models and active contours implemented in a level set framework. The shape detection is achieved by minimizing the Gibbs energy of the posterior probability function. The statistical shape model is built as a result of a learning process based on nonparametric probability estimation in a PCA reduced feature space formed by the Legendre moments of training silhouette images. The proposed energy is minimized by iteratively evolving an implicit active contour in the image space and subsequent constrained optimization of the evolved shape in the reduced shape feature space. Experimental results are also presented to show that the proposed method has very robust performances for images with a large amount of noise.

Keywords: Active contour, Legendre moments, statistical model, segmentation, shape detection.

1 Introduction

Active contour models have achieved enormous success in image segmentation and although there are number of ways to construct an active contour the most common approach is based on minimizing a segmentation functional. Construction of a prior shape constraint into the segmentation functional has recently become the focus of intensive research [1,2,3,4]. The early work on this problem has been based on principal component analysis (PCA) calculated for landmarks selected for a training set of shapes which are assumed to be representative of the shape variations. Tsai *et al.* [5] proposed a method to directly search solution in the shape space which is built by the signed

* This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/D077540/1].

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 51–58, 2011.

distance functions of aligned training images and reduced by PCA. In [6], Fussenegger *et al.* authors apply a robust and incremental PCA in order to improve segmentation results. Recently, it has been proposed to construct nonparametric shape prior by extending the Parzen density estimator to the space of shapes [7,8,9].

Foulonneau *et al.* [10] proposed an alternative approach for shape prior integration in the framework of parametric snakes. They proposed to define a $geometric$ shape prior based on a description of the target object shape using Legendre moments. A new shape energy term is defined as the distance between moments calculated for the evolving active contour and the moments calculated for a fixed reference shape priors. The main drawbacks of such an approach lies in its strong dependence on the shape alphabet used as reference. Indeed, as stated by the authors themselves in [10], this method is more related to *template matching* than to *shape learning*.

Inspired by the aforementioned results and especially by the approach proposed by Foulonneau *et al.* , the method proposed in this paper optimizes, within the level sets framework, model consisting of a prior shape probability model and image likelihood function conditioned on shapes. The statistical shape model results from a learning process based on nonparametric estimation of the posterior probability, in a low dimensional shape space of Legendre moments built from training silhouette images. Such approach tends to combine most of the advantages of the aforementioned methods, that is to say, it can handle multi-modal shape distributions, preserve a consistent framework for shape modeling and is free from any explicit shape distribution model.

The structure of this paper is as follows: The statistical shape model constructed in the space of the Legendre moments is explained in section 2.1; The level set active contour framework used in the proposed method is briefly explained in section 2.2; Section 2.3 defines the energy minimization problem, whereas in section 2.4 the proposed strategy for its minimization is explained in detail; Section 3 demonstrate the performance of the proposed method on images corrupted by severe random and structural noise; The conclusions are given in section 4.

2 Theory

The proposed method can be seen as constrained contour evolution, with the evolution driven by an iterative optimization of the posterior probability model that combines a prior shape probability and an image likelihood function. In this section all the elements of the proposed model along with the proposed optimization procedure are described in detail.

2.1 Statistical Shape Model of Legendre Moments

The method proposed in this paper, similarly to the method described in [10], uses shapes descriptors encoded by central-normalized Legendre moments $\boldsymbol{\lambda} = \{\lambda_{pq}, p + q \leq N_o\}$ of order N_o where p and q are non-negative integers, and therefore $\boldsymbol{\lambda} \in \mathbf{R}^{N_f}$ with $N_f = (N_o + 1)(N_o + 2)/2$. In the first instance, the mean vector $\bar{\boldsymbol{\lambda}}$ and the $N_f \times N_f$ covariance matrix $\mathbf{Q}$ are estimated for the central-normalized Legendre moments $\{\boldsymbol{\lambda}_i\}_{i=1}^{N_s}$ calculated for the shapes $\{\Omega_i\}_{i=1}^{N_s}$ from the training database. Subsequently the $N_f \times N_c$ projection matrix $\mathbf{P}$ is formed by the eigenvectors of the covariance matrix

$\mathbf{Q}$ that correspond to the largest N_c ($N_c \leq \min\{N_s, N_f\}$) eigenvalues. The projection of feature vectors $\{\boldsymbol{\lambda}_i\}_{i=1}^{N_s}$ onto the shape space, spanned by the selected eigenvectors, forms the feature vectors $\{\boldsymbol{\lambda}_{r,i}\}_{i=1}^{N_s}$:

$$\boldsymbol{\lambda}_{r,i} = \mathbf{P}^T(\boldsymbol{\lambda}_i - \bar{\boldsymbol{\lambda}}) \tag{1}$$

Finally the density estimation $P(\boldsymbol{\lambda}_r)$, with $\boldsymbol{\lambda}_r$ defined in the shape space, is performed up to a scale using $\boldsymbol{\lambda}_{r,i}$ as samples from the population of shapes and with the isotropic Gaussian function as the Parzen window:

$$P(\boldsymbol{\lambda}_r) = \sum_{i=1}^{N_s} \mathcal{N}(\boldsymbol{\lambda}_r; \boldsymbol{\lambda}_{r,i}, \sigma^2) \tag{2}$$

where $\mathcal{N}(\boldsymbol{\lambda}_r; \boldsymbol{\lambda}_{r,i}, \sigma^2) = \exp(-||\boldsymbol{\lambda}_r - \boldsymbol{\lambda}_{r,i}||^2/2\sigma^2)$

2.2 Level Set Active Contour Model

To detect and segment shapes a mechanism for taking into consideration the evidence about shape, present in an observed image, needs to be included. In this paper the region competition scheme of Chan-Vese is used for this purposes, with the energy given by:

$$E_{cv}(\Omega, \mu_\Omega, \mu_{\Omega^c}|I) = \int_\Omega (I - \mu_\Omega)^2\, dxdy + \int_{\Omega^c} (I - \mu_{\Omega^c})^2\, dxdy + \gamma|\partial\Omega| \tag{3}$$

where Ω^c represents the complement of Ω in the image domain and $|\partial\Omega|$ represent the length of the boundary $\partial\Omega$ of the region Ω. The above defined energy minimization problem can be equivalently expressed as maximization of the likelihood function:

$$P(I|\Omega) \propto \exp(-E_{cv}(\Omega, \mu_\Omega, \mu_{\Omega^c}|I)) \tag{4}$$

where $P(I|\Omega)$ could also be interpreted as a probability of observing image I when shape Ω is assumed to be present in the image. Introducing level set (embedding) function ϕ such that the Ω can be expressed in terms of ϕ as $\Omega = \{(x, y) : \phi(x, y) \geq 0\}$, as well as $\Omega^c = \{(x, y) : \phi(x, y) < 0\}$ and $\partial\Omega = \{(x, y) : \phi(x, y) = 0\}$. It can be shown that energy function defined in Eq.(3) is minimized by function ϕ given as a solution of the following PDE equation:

$$\frac{\partial\phi}{\partial t} = \left((I - \mu_{\Omega^c})^2 - (I - \mu_\Omega)^2\right)|\nabla\phi| + \gamma\nabla\left(\frac{\nabla\phi}{|\nabla\phi|}\right)|\nabla\phi| \tag{5}$$

with μ_Ω and μ_{Ω^c} representing respectively the average intensities inside and outside the evolving curve.

2.3 Energy Function

Introduced in the previous two sections distributions representing shape prior information and image intensity can be combined using Bayes rule and the Gibbs distribution model, resulting in the following energy function:

$$E(\boldsymbol{\lambda}_r) = E_{prior}(\boldsymbol{\lambda}_r) + E_{image}(\boldsymbol{\lambda}_r) \tag{6}$$

where the shape prior term is defined as:

$$E_{prior}(\boldsymbol{\lambda}_r) = -\ln\left(\sum_{i=1}^{N_s} \mathcal{N}(\boldsymbol{\lambda}_r; \boldsymbol{\lambda}_{r,i}, \sigma^2)\right) \tag{7}$$

and is built based on the shape samples Ω_i as explained in section 2.1. The image term is defined as:

$$E_{image}(\boldsymbol{\lambda}_r) = E_{cv}(\Omega, \mu_\Omega, \mu_{\Omega^c}|I)|_{\Omega=\Omega(\boldsymbol{\lambda}_r)} \tag{8}$$

where optimization of E_{cv} is constraint to shapes Ω from the estimated shape space $\Omega = \Omega(\boldsymbol{\lambda}_r)$ where $\Omega(\boldsymbol{\lambda}_r)$ denotes a shape from the shape space represented by the Legendre moments $\boldsymbol{\lambda} = \mathbf{P}\boldsymbol{\lambda}_r + \bar{\boldsymbol{\lambda}}$. The details of the optimization procedure for such energy are given in the next section.

2.4 Optimization

The proposed optimization procedure for minimization of the energy given in Eq.(6) is summarized in the following steps:

- Evolution of Ω according to Eq.(5):

$$\Omega^{(k)} \rightarrow \Omega^{'(k)} \tag{9}$$

  shape $\Omega^{(k)}$, from the previous algorithm iteration, is used as the initial shape and $\Omega^{'(k)}$ is the result of shape evolution. In the current implementation just single evolution iteration is used;
- Projection of the evolved shape into the shape space:

$$\Omega^{'(k)} \rightarrow \boldsymbol{\lambda}_r^{(k)} \tag{10}$$

  where $\boldsymbol{\lambda}_r^{(k)} = \mathbf{P}^T(\boldsymbol{\lambda}^{(k)} - \bar{\boldsymbol{\lambda}})$, and the central-normalized Legendre (c-nL) moments in vector $\boldsymbol{\lambda}^{(k)}$ are calculated using (L_{pq} are the 2D c-nL polynomials):

$$\lambda_{pq}^{(k)} = \frac{1}{|\Omega'^{(k)}|} \int_{\Omega'^{(k)}} L_{pq}\left(x, y, \Omega^{'(k)}\right) dxdy \tag{11}$$

- Shape space vector update:

$$\boldsymbol{\lambda}_r^{(k)} \rightarrow \boldsymbol{\lambda}_r^{'(k)} \tag{12}$$

  This step reduces the value of E_{prior} by moving $\boldsymbol{\lambda}_r^{(k)}$ in the steepest descent direction:

$$\boldsymbol{\lambda}_r^{'(k)} = \boldsymbol{\lambda}_r^{(k)} - \beta \left.\frac{\partial E_{prior}}{\partial \boldsymbol{\lambda}_r}\right|_{\boldsymbol{\lambda}_r = \boldsymbol{\lambda}_r^{(k)}} \tag{13}$$

  where

$$\frac{\partial E_{prior}}{\partial \boldsymbol{\lambda}_r} = \frac{1}{2\sigma^2} \sum_{i=1}^{N_s} w_i(\boldsymbol{\lambda}_r - \boldsymbol{\lambda}_{r,i}) \tag{14}$$

  with

$$w_i = \frac{\mathcal{N}(\boldsymbol{\lambda}_r; \boldsymbol{\lambda}_{r,i}, \sigma^2)}{\sum_{k=1}^{N_s} \mathcal{N}(\boldsymbol{\lambda}_r; \boldsymbol{\lambda}_{r,k}, \sigma^2)} \tag{15}$$

- Shape reconstruction from Legendre moments:

$$\boldsymbol{\lambda}_r^{\prime(k)} \rightarrow \Omega^{(k+1)} \tag{16}$$

where shape $\Omega^{(k+1)}$ is reconstructed using:

$$\Omega^{(k+1)} = \left\{ (x,y) : \left(\sum_{p,q}^{p+q \leq N_o} \lambda_{pq}^{\prime(k)} L_{pq}\left(x, y, \Omega^{\prime(k)}\right) \right) > 0.5 \right\} \tag{17}$$

with the Legendre moments $\lambda_{pq}^{\prime(k)}$ in vector $\boldsymbol{\lambda}^{\prime(k)}$ calculated from the shape space vector $\boldsymbol{\lambda}_r^{\prime(k)}$ using: $\boldsymbol{\lambda}^{\prime(k)} = \mathbf{P}\boldsymbol{\lambda}_r^{\prime(k)} + \bar{\boldsymbol{\lambda}}$

These steps are iterated until no shape change occurs in two consecutive iterations: $\Omega^{(k+1)} = \Omega^{(k)}$.

It should be pointed out that, unlike derivative based optimization methods such as [10], the shape descriptors need *not* be differentiable in the proposed method.

3 Experimental Results

A first set of experiments was carried out using a chicken image database consisting of 20 binary silhouette images with different sizes where 19 of them were used as training shapes for building the statistical prior model and the remaining image was used for testing (see Figure 1). The test images used for the method evaluation are shown in Figure 2. These images where obtained from a binary image by applying three different types of noise, namely, additive white Gaussian noise, structural noise for the simulation of occlusion and defects, as well as a combination of Gaussian and structural noise (hybrid noise). For all the results shown for this data, the same initial contour and the same key parameters $N_o = 40$ and $N_c = 10$ were used. For the test image with Gaussian noise the noise level is so high that even people with prior knowledge of the shape have difficulty in locating it visually. The segmentation result using the Chan-Vese model, which is well-known for its robustness to Gaussian noise, is shown in Figure 3(d). Figure 3(g) shows the segmentation result using the multi-reference method from [10],

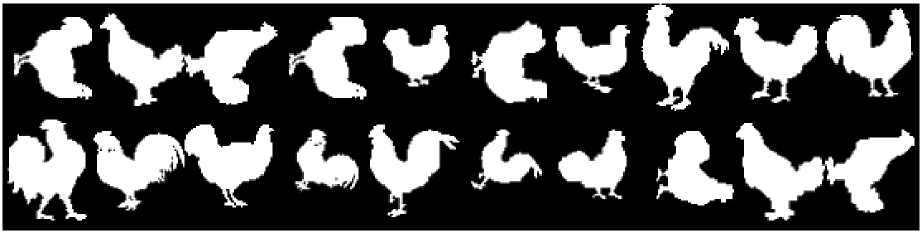


Fig. 1. The chicken image database where 19 images are used to build the statistical shape model and the remaining image (second from right in the bottom row) is used for testing

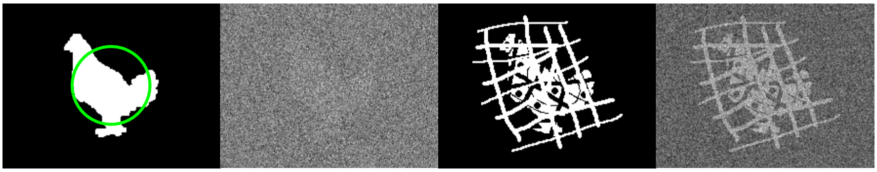


Fig. 2. Test images used for the evaluation of the proposed method. From left to right (i) Original noise-free binary test image with initial active contour shown as a circle at the center of the image; (ii) Image with severe white Gaussian noise; (iii) Image with structural noise; (iv) Image with hybrid noise.

where all the 20 training shapes were used as references. In this case, a range of different values of the method's design parameters (weights) were tried, but none of them ensured the algorithm convergence to the right result. Much better result was achieved using the proposed method as shown in Figure 3(a). As expected, the resulting shape living in the reduced feature space tends to have more regular appearance. The segmentation/detection results for the image with a large amount of structural noise are illustrated in Figure 3(b,e,h), where the necessity of shape prior constraint is clearly seen. Chan-Vese model without shape constraint completely failed by following the false structures. Although by increasing the weight associated with the length term (γ in Eq.(3)) the algorithm can avoid some of the false structures, it cannot properly locate the desired shape. Again, the multi-reference method failed to converge to the right result. Figure 3(c,f,i) show the results obtained for an image with both Gaussian and structural noise. As before Chan-Vese and multi-reference methods failed to recover original shape whereas the proposed method was able to detect the shape reasonably well. Although the main objective of the described experiment was to demonstrate a superior robustness of the proposed methods with respect to severe random and structural noise, the accuracy of the method was also tested on repeated experiments with different combination of the target image and structural noise pattern. It transpired that the proposed method was able to localize object boundary with an average accuracy of 1.2, 1.7 and 2 pixels when operating respectively on images with Gaussian, structural and hybrid noise.

A second set of experiments was carried out using gray scale images. An example of a test image used in this experiment is shown in Figure 5 where the objective was to segment the cup. To build the shape space for the "cup objects" an image set composed of 20 binary cup silhouette images (shown in Figure 4), from the *MPEG7 CE shape-1 Part B* database, was used. It can be clearly seen that the training shapes integrate a large shape variability. Results of segmentation using the Chan-Vese, multi-reference and the proposed method are shown in Figure 5. Assuming that the goal of the segmentation was to recover the shape of the cup, the proposed method leads to much more accurate result with the final shape segmentation not altered by the drawing on the cup or by books and a pen in the background. This clearly demonstrates that, the proposed method is much more robust than the other two tested methods with respect to "shape distractions" present in the data. The final result can be seen as a good compromise between image information and the prior shape constraints imposed by the training data set used.

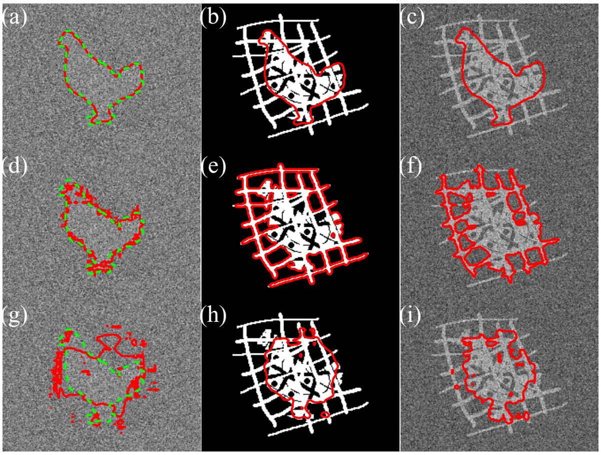


Fig. 3. Results for the test data, shown in Figure 2, obtained for: proposed method (a-c); Chan-Vese method (d-f); multi-reference method proposed in [10] (g-i). The red solid curves depict segmentation results, whereas the desired results (plotted for the images with the Gaussian noise only) are shown as green dash lines.

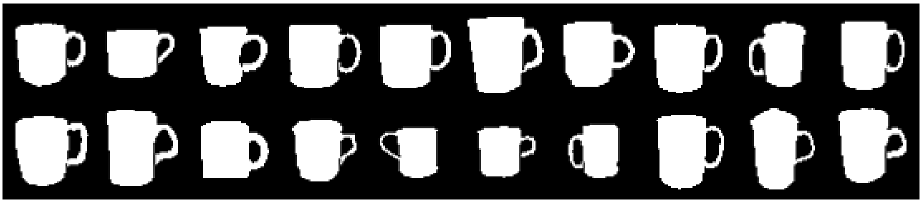


Fig. 4. The cup image set used to build the shape space

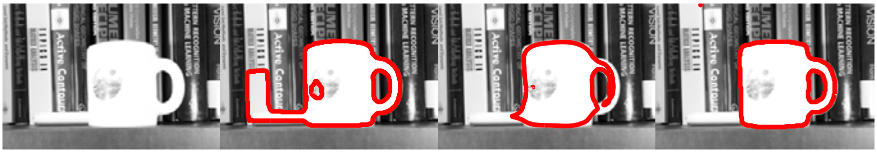


Fig. 5. From left to right (i) an image to be segmented, (ii) result of segmentation using Chan-Vese model, (iii) result of the segmentation using the multi-reference method from [10], (iv) result of the segmentation using the proposed method

4 Conclusions

The paper describes a novel method for shape detection and image segmentation. The proposed method can be seen as constrained contour evolution, with the evolution driven by an iterative optimization of the posterior probability function that combines a prior shape probability and the image likelihood function. The prior shape probability function is defined on the subspace of Legendre moments and is estimated, using Parzen window method, on the training shape samples given in the estimated beforehand shape space. The likelihood function is constructed from conditional image probability distribution, with the image modelled to have regions of approximately constant intensities, and regions defined by the shape which is assumed to belong to the estimated shape space. The resulting constrained optimization problem is solved using combinations of level set active contour evolution in the image space and steepest descent iterations in the shape space. The decoupling of the optimization processes into image and shape spaces provides an extremely flexible optimization framework for general statistical shape based active contour where evolution function, statistical model, shape representation all become configurable. The presented experimental results demonstrate very strong resilience of the proposed method to the random as well as structural noise present in the image.

References

1. Houhou, N., Lemkaddem, A., Duay, V., Allal, Abdelkarim, Thiran, J.-P.: Shape prior based on statistical MAP for active contour segmentation. In: ICIP (2008)
2. Kim, J., Çetin, M., Willsky, A.S.: Nonparametric shape priors for active contour-based image segmentation. Signal Process. 87(12), 3021–3044 (2007)
3. Lecellier, F., Jehan-Besson, S., Fadili, J., Aubert, G., Revenu, M., Saloux, E.: Region-based active contour with noise and shape priors. In: ICIP, pp. 1649–1652 (2006)
4. Erdem, E., Tari, S., Vese, L.A.: Segmentation using the edge strength function as a shape prior within a local deformation mode. In: ICIP, pp. 2989–2992 (2009)
5. Tsai, A., Yezzi, A., Wells, W.M., Tempany, C., Tucker, D., Fan, A., Grimson, W., Willsky, A.: A shape-based approach to the segmentation of medical imagery using level sets. IEEE Trans. Med. Imaging 22(2), 137–154 (2003)
6. Fussenegger, M., Roth, P., Bischof, H., Deriche, R., Pinz, A.: A level set framework using a new incremental, robust Active Shape Model for object segmentation and tracking. Image Vision Comput. 27(8), 1157–1168 (2009)
7. Cremers, D., Osher, S.J., Soatto, S.: Kernel density estimation and intrinsic alignment for shape priors in level set segmentation. Int. J. Comput. Vision 69(3), 335–351 (2006)
8. Rousson, M., Cremers, D.: Efficient kernel density estimation of shape and intensity priors for level set segmentation. In: Duncan, J.S., Gerig, G. (eds.) MICCAI 2005. LNCS, vol. 3750, pp. 757–764. Springer, Heidelberg (2005)
9. Rousson, M., Paragios, N.: Prior knowledge, level set representation and visual grouping. Int. J. Comput. Vision 76(3), 231–243 (2008)
10. Foulonneau, A., Charbonnier, P., Heitz, F.: Multi-reference shape priors for active contours. Int. J. Comput. Vision 81(1), 68–81 (2009)

Efficient Image Segmentation Using Weighted Pseudo-Elastica

Matthias Krueger, Patrice Delmas, and Georgy Gimel'farb

Department of Computer Science,
The University of Auckland
m.krueger74@gmail.com

Abstract. We introduce a new segmentation method based on second-order energies. Compared to the related works it has the significantly lower computational complexity $O(N \log N)$. The increased efficiency is achieved by integrating curvature approximation into a new bidirectional search scheme. Some heuristics are applied in the algorithm at the cost of exact energy minimisation. Our novel pseudo-elastica core algorithm is then incorporated into a user-guided segmentation scheme which represents a generalisation of classic first-order path-based schemes to second-order energies while maintaining the same low complexity. Our results suggest that, compared to first-order approaches, it scores similar or better results and usually requires considerably less user-input. As opposed to a recently introduced efficient second-order scheme, both closed contours and open contours with fixed endpoints can be computed with our technique.

Keywords: Image segmentation, second-order energy, curvature regularity, active contour.

1 Introduction

The active contour (AC) introduced by Kass et al. [1] is a successful and seminal concept in Computer Vision. The main idea is to capture an object of interest by an evolving contour which converges towards the boundary of the object. The evolution is guided by an internal force, which imposes regularity on the contour, and an external force attracting the contour to image features. Main drawback of the original AC and its main successor, the geodesic active contour (GAC) [2], is that the respective minimised energies are not convex and curve evolution due to gradient descent usually converges to a steady state that is locally rather than globally optimal. During the last decade several techniques have been introduced to overcome this drawback. Cohen and Kimmel [3] applied Sethian's fast marching method [4] to efficiently compute globally optimal GACs with given endpoints. Later, Boykov and Kolmogorov [5] used graph cuts to obtain closed globally optimal GACs, and Appleton and Talbot [6] achieved the same goal by skilful application of Cohen and Kimmel's [3] scheme in a curved product space.

The above mentioned approaches [3,5,6] have in common that they base on first-order energies, i.e. suitably weighted arc length functionals. Subsequently, it has shown that first-order techniques, even scale invariant ones (see [7]), have a bias toward short curves. Therefore, segmentation energies including second-order, i.e. curvature-, terms

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 59–67, 2011.

have recently attracted considerable interest. Schoenemann and Cremers [7] find global minima of a ratio energy including curvature terms. However, large product graphs are required for this method, resulting in computing times of minutes and hours rather than seconds. Later, Schoenemann et al. [8] proposed a similar technique for region-based image segmentation that also suffered from high running times. Recently, Zehiry and Grady introduced efficient algorithms with curvature regularity for both region- and edge-based energies ([9] and [10], respectively), transforming the original problem to an graph cut optimisation problem. The authors report running times of a few seconds.

In this paper we introduce a new approach to curvature regularised image segmentation. Its core is a novel bidirectional Dijkstra search scheme in a graph-based framework. Due to its computational complexity of $O(N \log N)$, the core algorithm can be seen as a generalisation of first-order shortest path algorithms such as [11,4] and related schemes for image segmentation [3]. The running times are typically below one second. Yet heuristics have to be applied to achieve such computational efficiency, therefore the computed contours are only approximately globally optimal. In contrast to the approach in [10] our algorithm also allows for the computation of open contours with fixed endpoints, which is e.g. beneficial for feature segmentation from face- or medical data.

The paper is organised as follows: Two new segmentation energies are introduced in Section 2. Algorithms are proposed in Section 3, that compute approximately globally minimal open curves. The core algorithm is integrated into a user-guided schemes for image segmentation in Section 4. Section 5 presents results on several images, including quantitative comparisons with the state-of-the-art techniques. The conclusions in Section 6 complete the paper.

2 Second-Order Energies for Active Contours

Let $I : \Omega \to \mathbb{R}^+$ be a greyscale image on the rectangular domain Ω and $\Gamma : J \to \Omega$ be a closed contour with $J = [0, 1]$. Further, let $f : \Omega \to \mathbb{R}^+$ denote an edge indicator function, taking small values close to desired image features and larger values elsewhere. We propose the following two segmentation energies:

$$E_1(\Gamma) = \int_\Gamma f(s)(c\kappa^2 + 1)\,\mathrm{d}s \quad \text{and} \quad E_2(\Gamma) = \int_\Gamma \left(c\kappa^2 + f(s)\right)\,\mathrm{d}s\ , \tag{1}$$

where κ is the curvature of Γ, and $c > 0$ is a constant. There is a close relationship between the energies E_1 and E_2 proposed here and both, the classical GAC energy (E_1 with $c = 0$, see [2]) and the elastica energy (E_1 with $f \equiv 1$, see e.g. [9]). Thus, energies E_1 and E_2 can equally be interpreted as second-order regularisations of the GAC functional and weighted versions of the elastica, attracted to certain image features.

The key difference between the two proposed energies is the following: while the curvature term is weighted with the feature detector function f in E_1, it has a constant weight c in E_2. Hence, the curvature regularisation prior is strong in E_2, making this energy suitable for robust segmentation of objects with comparatively simple shapes, such as approximately round or elliptic objects, even under strong noise. In energy E_1, the curvature regularisation is relaxed when the contour proceeds along an edge, therefore this energy can model a broader range of object shapes.

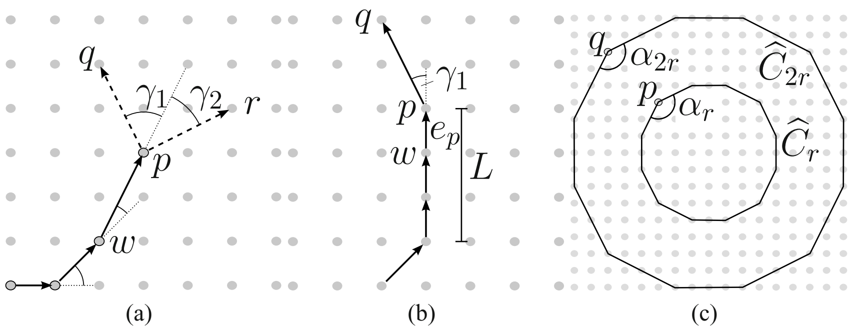


Fig. 1. The idea of dynamic curvature estimation (a). Refined curvature estimation is achieved by tracking straight segments over multiple edges (b). Representation of two discrete circles (c).

3 Efficient Approximation of Global Minimisers

3.1 Open Contours with Directional Constraints in One Endpoint

The graphs used here are directed 2D grid graphs $G = \langle \mathcal{V}, \mathcal{E} \rangle$ with vertices $\mathcal{V}$ that are embedded into $\mathbb{R}^2$ in a regular grid-like manner. The edges are generated by topologically identical neighbourhood systems $\mathcal{D}$ for all vertices. To obtain realistic approximations of the curvature, a large 32-neighbourhood system is chosen (cf. [7]).

Core of our algorithm is a Dijkstra scheme [11] starting from one point (possibly with directional constraints) and terminating once the endpoint is reached. The novelty is that curvature is integrated and the edge weights are computed dynamically. Note, that more details of the algorithm, including pseudo-code, can be found in [12].

Dynamically computing curvature. The estimation of the curvature is achieved similarly as in e.g. [7]: by computing the angle between two adjacent edges. However, to obtain an efficient algorithm, we incorporate curvature estimation into a greedy algorithm. Figure 1(a) illustrates the main idea of approximating the curvature, under the assumption that the pseudo-minimal path up to pixel p has already been computed. In order to update the neighbours of p, the edge weights are defined dynamically: the curvature is approximated analysing the respective angles in the contour (γ_1 and γ_2 in Fig. 1(a)). Apparently, the curvature cannot be defined for a single edge, but rather for a pair of edges. Thus, the weight of an edge is computed from the curvature and a weighted average of f on the edge e and its predecessing edge e_p (see (3) below). For the energy E_1, the weight ω of the edge $e = (p, q)$ is computed as follows (with $e_p = (w, p)$):

$$\omega(e) = \widehat{f}(e_p, e)(c\kappa_{\gamma_1}^2 + 1)\ell(e) \tag{2}$$

with $$\widehat{f}(e_p, e) = \ell(e_p)/\big(\ell(e_p) + \ell(e)\big) \cdot f(e_p) + \ell(e)/\big(\ell(e_p) + \ell(e)\big) \cdot f(e) \tag{3}$$

with $\ell(.)$ the length of an edge. E_2 is treated analogously. For the starting edge we assume $\kappa = 0$ and define $\omega(e) = f(e)\ell(e)$.

The definition of the curvature term κ_{γ_1} is discussed next. A first guess is $\kappa_{\gamma_1} = C \cdot \gamma_1$ with C a suitable weight depending on the length of the adjacent edges (as in [7]).

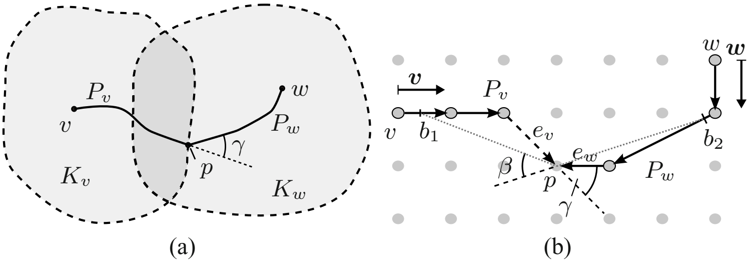


Fig. 2. The bidirectional Dijkstra scheme: (a) illustrates the key idea, and (b) details the notation

Yet considerations regarding the scaling behaviour of the curvature (detailed in [12]) suggest that this approach has to be enhanced to suit the proposed framework (see Fig. 1(b)): the algorithm tracks along the minimal path, when a straight line is extended over multiple edges, and sums up the total length L of the straight segment. If the absolute curvature angle $|\gamma_1|$ does not exceed a certain threshold $p_{\kappa,1}$ (roughly $2 \cdot \frac{2\pi}{|\mathcal{D}|}$), the estimated curvature in p is obtained by dividing the curvature angle γ_1 by L. Otherwise, γ_1 is divided by the length of the incoming edge $e_p = (w, p)$. By this means, we achieve that the curvature scales as desired for sufficiently smooth – in a discrete sense – contours differing only in their scale (cf. Fig. 1(c)).

3.2 Open Contours with Directional Constraints in Both Endpoints

The bidirectional Dijkstra approach (Fig. 2). Starting from the endpoints v and w, the distance functions d_v and d_w are computed simultaneously – each by a dynamic Dijkstra scheme as described in Sect. 3.1. The sets of vertices labeled 'known' or 'trial' by the Dijkstra schemes D_v and D_w after i iterations are denoted by $K_v(i)$ and $K_w(i)$ (Fig. 2(a)). With increasing i, the sets $K_v(i)$ and $K_w(i)$ propagate and, eventually, start to overlap (Fig. 2(a)). For a vertex p in the intersection the preliminary pseudo-minimal paths P_v and P_w (as in Section 3.1) can be tracked back to the endpoints v and w, respectively. We measure the smoothness of the intersection between the partial paths in vertex p as follows (see Fig. 2(b)): first, the local curvature is computed from the angle γ, describing the change of direction of the adjacent edges e_v and e_w. If $|\gamma|$ is below a certain threshold $p_{\kappa,2}$ (e.g. $\pi/4$) the intersection is at least locally sufficiently smooth. In this case, the curvature is reestimated in a larger neighbourhood in order to impose smoothness on a larger scale. The simple scheme in [13] is used: given a neighbourhood size parameter m , the scheme localises the two points b_1, b_2 that have the (arc length-) distance m from p. Then, the refined curvature is computed from the change of direction β in the m-neighbourhood (Fig. 2(b)). A score for the concatenated path is summed up from the partial distance values $d_v(p)$ and $d_w(p)$, and the intersection energy. If a score is lower than the current minimum, the score and the respective intersection point are stored. The algorithm terminates as soon as the sum of the partial distances on the propagating fronts exceeds the hitherto minimal score. The pseudo-minimal path is then tracked back from the stored intersection point. Note that the idea of bidirectional

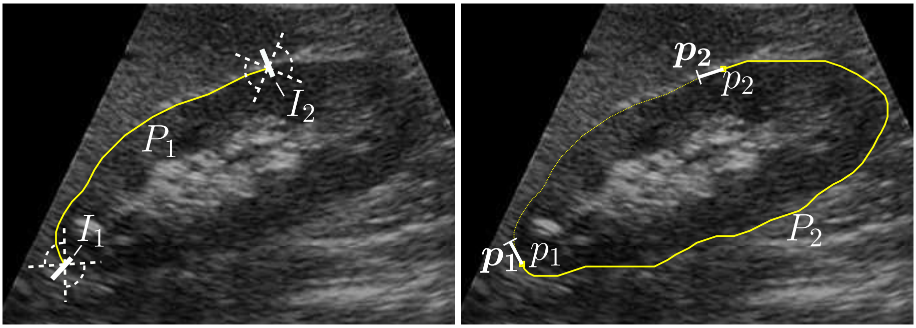


Fig. 3. Example for user-guided segmentation (ultrasound kidney image, energy E_2 with $c = 1$): two lines I_i across the object boundary are provided by the user. Only rough constraints are applied for the computation of P_1 (left). Section P_2 is constrained to smoothly close P_1 (right).

search goes back to Pohl [14]. The novelty of our algorithm lies in the generalisation of this notion to second-order energies.

For an image with N pixels and neighbourhood size $|\mathcal{D}|$ the graphs in our method have N vertices and about $|\mathcal{D}| \cdot N$ edges. Assuming $|\mathcal{D}|$ to be constant, this results in an asymptotic complexity of $O(N \log N)$ for the algorithms proposed above.

4 Application to Image Segmentation

For the experiments shown in this paper we use the following edge detector function:

$$f^{\pm}(e) = 1/(1 + (\max\{0, \pm\langle \nu, \overline{\nabla I_\sigma(e)}\rangle\})^2) \ . \tag{4}$$

Here, I_σ is an image smoothed by convolution with a suitable kernel, $\overline{\nabla I_\sigma}$ denotes the gradient of I_σ averaged along an edge e and ν is the unit normal of e. Depending on whether the object or the background is brighter, f^+ or f^- has to applied. In the following, these two complementary functions are subsumed under the notation f.

Algorithm for closed contours. Apart from the obvious application of the proposed algorithm to open contours, it can also be used to compute piecewise pseudo-optimal closed contours. As opposed to related first-order schemes (e.g. [3]), smooth transitions are ensured by the directional constraints in the intersection points.

The user has to provide n ordered streaks (e.g. lines) $I_1, \ldots, I_n$ across the boundary of the object of interest. The robustness and efficiency of the algorithm is further improved, when the streaks are assumed to run roughly perpendicular through the boundary (cf. Fig. 3). For each section, the above algorithm for open contours is applied to find a pseudo-minimal path satisfying the following constraints:

- A contour connecting the sets I_1 and I_2 is computed. f is tested with both possible orientations, and the path with the lower energy indicates the proper orientation;
- For the interior sections ($P_i, 1 < i < n$) the path is constrained to start from the second-last vertex of the previous section (P_{i-1}) in direction of the last edge of P_{i-1}. The conditions in the endpoint are as described for the first section above;

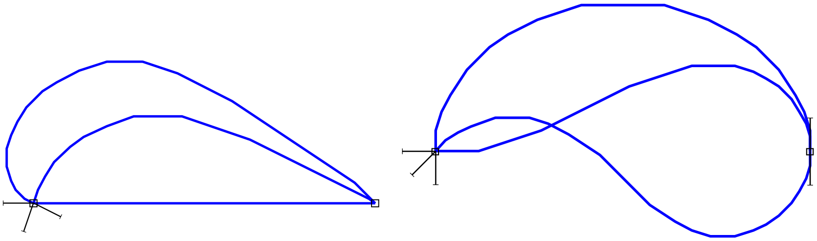


Fig. 4. Pseudo-elastica (classic case, $f \equiv 1$ in (1)) connecting two points with directional constraints in the starting point only (left, $c = 30$) and both endpoints (right, $c = 30$)

– For section P_n, the start direction is derived from P_{n-1}, and the ending direction is inferred from P_1 (Fig. 3 (right)).

5 Experimental Results

We compared the pseudo-elastica to state-of-the-art segmentation methods with respect to accuracy and efficiency. All algorithms were coded in C++, and the experiments were run on an Intel Core 2 Duo 2.5 GHz machine. The parameters inside the pseudo-elastica core algorithms were kept constant across all experiments: $p_{\kappa,1} = 0.15\pi, p_{\kappa,2} = \frac{\pi}{4}$, $m = 7$, see [12]. First we validated our approach by simulating the case of the classic elastica ($f \equiv 1$ in either energy in (1)). The results (see Fig. 4) indicate that despite the use of heuristics our method delivers a good approximation of the exact elastica (cf. [15]). Subsequently, the pseudo-elastica were compared with two top recognised schemes in the field of medical imaging: the globally optimal AC [3] and the graph cut based GAC [5]. Note that a comparison with the algorithm proposed in [10] is the subject of a subsequent paper. We emphasise that in contrast to [10] our technique is also applicable to open contours with fixed endpoints (bottom row of Fig. 5).

Figure 5 shows sample results on images of different qualities and noise levels, and a quantitative evaluation is given in Table 1. Ground truth contours were delineated manually by a medical expert. Both, visual impression and the statistical values in Table 1, suggest that the pseudo-elastica algorithm performs similarly well or better than its competitors in many cases. Further, our experiments showed that considerably less user input is required by the pseudo-elastica method than Cohen and Kimmel's first-order technique [3]. For the corpus callosum image with its weak edges, for instance,

Table 1. Accuracy evaluation of the proposed algorithm

Image	Size	**Pseudo-Elastica**					**Cohen/Kimmel [3]**				**Graph cut [5]**			
	(in pixels)			error (pix.)		time		error (pix.)		time		error (pix.)		time
		Engy.	c	mean	max	in s	w	mean	max	in s	σ	mean	max	in s
C. Callosum	276 × 200	E_1	1.0	1.1	3.3	0.3	0.25	1.2	4.3	0.2	1	1.4	7.1	0.1
Head	367 × 303	E_2	1.0	1.7	4.3	0.5	0.1	2.1	11.9	0.6	2.5	5.9	17.8	0.3
Filament	322 × 298	E_1	10.0	1.2	4.3	0.2	0.025	2.5	12.7	0.3				

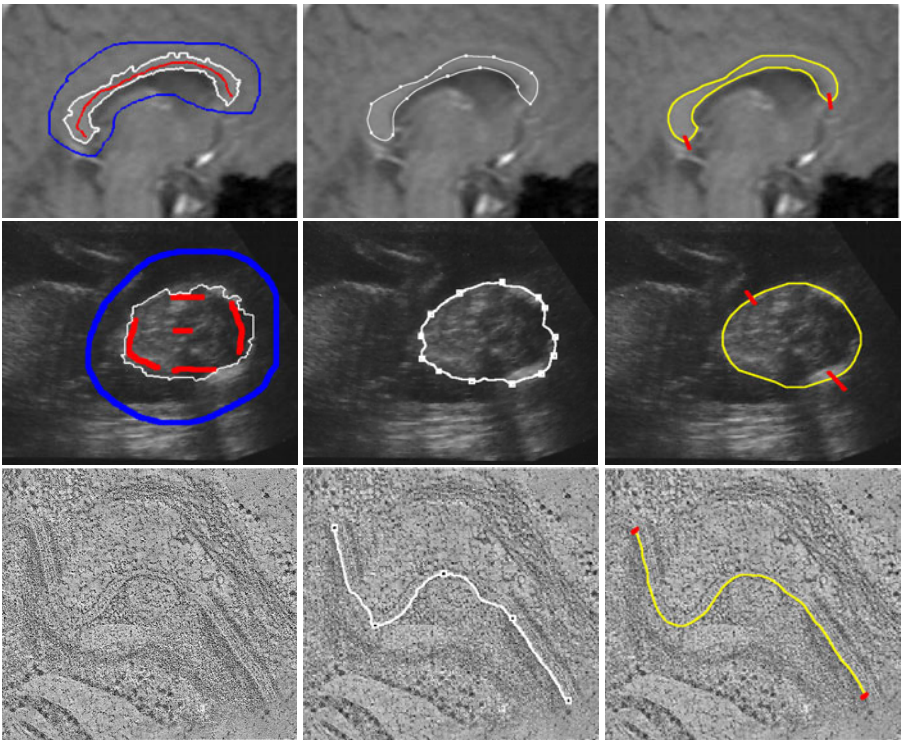


Fig. 5. Comparison of pseudo-elastica (right, with input streaks) with Cohen/Kimmel AC (middle, with nodes) and graph cut GAC (left, with foreground/background seeds)

more than 10 nodes had to be provided to Cohen and Kimmel's algorithm by the user. This is a tedious and time consuming task, that often requires several trials before an acceptable result is obtained. On the other hand, two lines across the object boundary are mostly sufficient to obtain a robust and accurate result with pseudo-elastica.

Second-order techniques such as the pseudo-elastica are particularly suited for the segmentation of ultrasound images. Such images usually contain a great amount of speckle noise and large gaps in the object boundaries, rendering them difficult to process. The graph cut method can hardly cope with these conditions (see the middle row in Fig. 5), and Cohen/Kimmel's method yields satisfactory results only if many nodes are provided by the user. The pseudo-elastica in conjunction with energy E_2 usually achieve a robust result with little user input. Despite the additional averaging of the detector function over two adjacent edges in (3) the accuracy of our method is similar or better than that of the compared schemes. Finally, our experiments support the fact that the computational complexity of the pseudo-elastica is the same as the one of Cohen and Kimmel's method [3]: $O(N \log N)$. All contours shown in Fig. 5 were computed in less than one second. We note that further experiments can be found in [12].

6 Conclusion

We have developed a novel framework for 2D image segmentation.Two segmentation energies penalising curvature were introduced, and a new algorithm was developed to efficiently approximate these minimisers using a bidirectional Dijkstra-like search. The core algorithm has the complexity[1] $O(N \log N)$, which is significantly less than existing globally optimal segmentation methods [7,16] incorporating curvature regularity.

Results on various images, including noisy ultrasound images, suggest that our method scores favourably against state-of-the-art algorithms [5,3]. Equally good or better results were obtained with significantly less user input. Due to typical processing times of less than a second, the presented method constitutes a feasible alternative to first-order shortest path approaches such as [3].

Main drawback of the pseudo-elastica compared to other methods such as [3,5,7,10] is that the computed contours are not exact energy minimisers. Heuristics are applied in order to obtain an efficient scheme for practical use, and some parameters inside the algorithm have to be set. Yet the algorithm showed robustness with respect to the choice of the parameters, since one set of values could be used across all experiments.

Like other shortest path-based techniques (e.g. [3,7]), pseudo-elastica cannot detect multiply-connected regions. This distinguishes the path-based methods from the implicit methods, such as level sets and graph cuts. Moreover, there is no obvious extension of the method to 3D volumetric images. Yet Ardon and Cohen [17] have applied shortest path techniques to 3D surface segmentation. Studying, whether pseudo-elastica can be embedded into a similar scheme is an exciting avenue for future research.

References

1. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. IJCV 1(4), 321–331 (1988)
2. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. IJCV 22(1), 61–79 (1997)
3. Cohen, L.D., Kimmel, R.: Global minimum for active contour models: A minimal path approach. IJCV 24(1), 57–78 (1997)
4. Sethian, J.: A fast marching level set method for monotonically advancing fronts. Proc. National Academy of Sciences 93, 1591–1595 (1996)
5. Boykov, Y., Kolmogorov, V.: Computing geodesics and minimal surfaces via graph cuts. In: Proc. ICCV 2003 (2003)
6. Appleton, B., Talbot, H.: Globally optimal geodesic active contours. Journ. of Math. Imag. and Vis. 23(1), 67–86 (2005)
7. Schoenemann, T., Cremers, D.: Introducing curvature into globally optimal image segmentation: Minimum ratio cycles on product graphs. In: Proc. ICCV 2007 (2007)
8. Schoenemann, T., Kahl, F., Cremers, D.: Curvature regularity for region-based image segmentation and inpainting: A linear programming relaxation. In: Proc. ICCV 2009 (2009)
9. El-Zehiry, N.Y., Grady, L.: Fast global optimization of curvature. In: Proc. CVPR 2010 (2010)
10. El-Zehiry, N.Y., Grady, L.: Optimization of weighted curvature for image segmentation (2010) (preprint), arXiv:1006.4175v1 [cs.CV]

[1] N denotes the number of image pixels.

11. Dijkstra, E.: A note on two problems in connexion with graphs. Num. Math. 1, 269–271 (1959)
12. Krueger, M.: Segmentation of Surfaces Using Active Contours. PhD thesis, The University of Auckland (February 2011)
13. Bennett, J.R., MacDonald, J.S.: On the measurement of curvature in a quantized environment. IEEE Transactions on Computers 24(8), 803–820 (1975)
14. Pohl, I.: Bi-directional search. Machine Intelligence 6, 127–140 (1971)
15. Mumford, D.: Elastica and computer vision. In: Bajaj, C. (ed.) Algebraic Geometry and its Applications, pp. 491–506. Springer, New York (1994)
16. Windheuser, T., Schoenemann, T., Cremers, D.: Beyond connecting the dots: A polynomial-time algorithm for segmentation and boundary estimation with imprecise user input. In: Proc. ICCV 2009 (2009)
17. Ardon, R., Cohen, L.D.: Fast constrained surface extraction by minimal paths. IJCV 69(1), 127–136 (2006)

Automatic Conic and Line Grouping for Calibration of Central Catadioptric Camera

Wenting Duan and Nigel M. Allinson

Department of Electrical and Electronic Engineering, University of Sheffield, U.K.,
School of Computer Science, University of Lincoln, U.K.
elq06wd@sheffield.ac.uk, nallinson@lincoln.ac.uk

Abstract. This paper presents an automatic method for detecting the conics from an omnidirectional image and use the conics for camera calibration. The method assumes the camera system is orientated in a way that the mirror axis is orthogonal to the floor. In this way two special cases of parallel lines can be obtained: one set of lines are orthogonal to the mirror axis; and the other set is lines that are coplanar with the mirror axis. Based on these special lines' geometric properties under central catadioptric projection, we show that automatic calibration of catadioptric from a single image can be achieved. The experiment results show that the method not only improve the accuracy of the conic fitting in the image but also provide robustness against conic occlusion and noise.

Keywords: Conic fitting, catadioptric camera, calibration, vanishing points.

1 Introduction

Accurate calibration parameters of camera are important in the application of 3D reconstruction, ego-motion, photogrammetry, etc. With the trend of using catadioptric systems in various applications where a wide field of view is required, various techniques for calibration of catadioptric cameras are devloped. A catadioptric camera can be calibrated using two or more images of the same scene, e.g. Kang [2] and Micusik [7]. Some researchers use planar grids. For example, Scaramuzza [4] described the catadioptric image projection with a Taylor series expansion. The coefficients of the expansion model are obtained by solving a least-square linear minimisation problem. In Mei and Gasparini's work [5][6], catadioptric homography is computed using images of planar grids. Factors such as misalignment of mirror and camera, camera-lens distortion are taken into account. Another route for calibrating catadioptric camera is to use geometric properties such as line projections from a single catadioptric image. Geyer and Daniilidis [11] proposed to use two sets of parallel lines for calibrating catadioptric camera. This method is designed for para-catadioptric cameras, the aspect ratio of the camera is assumed to be one, and then the parallel lines projected on the image plane are circular arcs with collinear centres. Ying [8] demonstrated the use of line and sphere projections in central catadioptric camera calibration. Ying's experiment show that the projection of sphere provide better conic fitting than line projection since sphere projection provide bigger portion of a conic. Hence a more accurate calibration is resulted. However, the requirement of

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 68–75, 2011.

at least three sphere projections in building environment is not easy to satisfy. Our proposed method also uses line projections in the catadioptric image. The idea is motivated by the fact that large amount of regular geometric structures such as parallel and orthogonal lines are normally presented in images of man-made objects. Research papers [1, 8-11] already show that geometric properties of conics (line projections) in the catadioptric image enable the calibration. If conics can be automatically detected, then the whole calibration process from a single catadioptric image can be made automatic. The problem lies at the automatic conic fitting. As pointed out by Barreto [1], under the central catadioptric projection, line in the 3D world is generally mapped into a small arc of a conic. This raises the difficulty of accurately estimating the conic parameters from the image itself.

This paper presents an automatic method for detecting the conics from an omnidirectional image and use the conics for camera calibration. The method assumes the camera system is orientated in a way that the mirror axis is orthogonal to the floor. The approach is based on the geometric properties of two special parallel lines sets: one set contain parallel line projections (i.e. conics) that are orthogonal to the mirror axis; and the other set include line projections that are coplanar with the mirror axis. The performance of the proposed method is evaluated on both synthetic and real data.

2 Method

The proposed method uses the unifying model of central catadioptric projection developed by Barreto [12] (shown in Figure 1(a)). A 3D world point $\boldsymbol{X}$ is projected on the unit sphere at $\boldsymbol{X_C}$. The unit sphere is centre at the focal point of mirror and denoted as $\boldsymbol{O}$. The unifying model represents this transformation by a 3×4 matrix $\boldsymbol{P}$

$$\boldsymbol{P} = \boldsymbol{R}[\boldsymbol{I}| - \boldsymbol{C}] \tag{1}$$

After computing $\boldsymbol{X_C} = \boldsymbol{PX}$, the point $\boldsymbol{X_C}$ is then mapped to the point $\bar{\boldsymbol{x}}$ in the sensor plane $\boldsymbol{\Pi}_\infty$. This transformation is modelled using function $\hbar$, in which the non-linearity of the mapping is contained. The point $\boldsymbol{O_C}$ with coordinates $(0, 0, -\xi)^T$ is the other projection centre which re-projects the point $\boldsymbol{X_C}$ on the unit sphere to $\bar{\boldsymbol{x}}$ in the sensor plane $\boldsymbol{\Pi}_\infty$. The function $\hbar$ is written as

$$\hbar(\boldsymbol{x}) = \begin{bmatrix} \frac{x}{\sqrt{x^2+y^2+z^2}} \\ \frac{y}{\sqrt{x^2+y^2+z^2}} \\ \frac{z}{\sqrt{x^2+y^2+z^2}} + \xi \end{bmatrix} \tag{2}$$

Finally, a collineation $\boldsymbol{H_C}$ is applied to transform $\bar{\boldsymbol{x}}$ to obtain the point $\hat{\boldsymbol{x}}$ in the catadioptric image plane, i.e. $\hat{\boldsymbol{x}} = \boldsymbol{H_C}\bar{\boldsymbol{x}}$. $\boldsymbol{H_C}$ is written as

$$\boldsymbol{H_C} = \boldsymbol{K_C R_C M_C} \text{ where } \boldsymbol{M_C} = \begin{bmatrix} \varphi - \xi & 0 & 0 \\ 0 & \xi - \varphi & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

$\boldsymbol{M_C}$ changes according to the mirror type and shape, $\boldsymbol{K_C}$ is the camera calibration matrix and $\boldsymbol{R_C}$ is the rotation matrix of the mirror relative to the camera. The majority of the catadioptric sensors commercially available have their mirror accurately aligned with the camera, i.e. the conventional camera is not rotated with relation to the mirror surface. Therefore, the rotation matrix $\boldsymbol{R_C} = \boldsymbol{I}$ and $\boldsymbol{H_C}$ is an affine transformation. The line at infinity $\hat{\boldsymbol{\pi}}_\infty$ is the intersection of the plane at infinity $\boldsymbol{\Pi}_\infty$ and the Catadioptric image plane $\widehat{\boldsymbol{\Pi}}$. Since $\boldsymbol{H_C}$ is affine, then $\bar{\boldsymbol{\pi}}_\infty = (0, 0, 1)^\mathrm{T}$ in $\boldsymbol{\Pi}_\infty$ is mapped to $\hat{\boldsymbol{\pi}}_\infty = (0, 0, 1)^\mathrm{T}$ in $\widehat{\boldsymbol{\Pi}}$. Generally, the calibration of a central catadioptric system is to obtain mirror parameter ξ and the collineation matrix $\boldsymbol{H_C}$. For parabolic mirror case, $\xi = 1$ and the calibration is much easier. The lines under the projection of parabolic sensor are mapped to circles in the image plane if aspect ratio of the camera is one and skew factor is zero [11]. Our target is then the more complicated hyperbolic case where ξ is unknown.

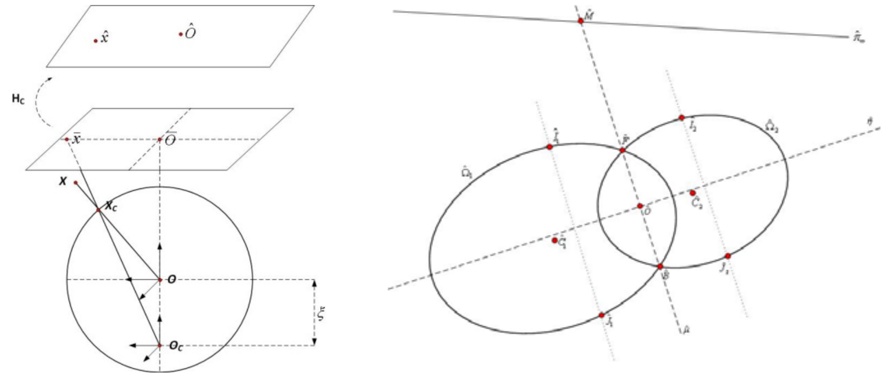
$$\boldsymbol{K_C} = \begin{bmatrix} r.f_e & s & u_0 \\ 0 & f_e & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

r is the aspect ratio, f_e is the effective focal length, s is the skew factor and $(u_0, v_0\ )$ is the principal point. To obtain ξ and $\boldsymbol{H_C}$ for a central catadioptric system from a single image, the key is the location of the absolute conic $\widehat{\boldsymbol{\Omega}}_\infty$ and the equation of cross ratio $\{\widehat{\boldsymbol{C}}, \widehat{\boldsymbol{N}}; \widehat{\boldsymbol{M}}, \widehat{\boldsymbol{O}}\}$. These are derived from

$$\widehat{\boldsymbol{\Omega}}_\infty = \boldsymbol{H_C^{-T}} \boldsymbol{H_C^{-1}} \tag{5}$$

$$\left\{\widehat{\boldsymbol{C}}, \widehat{\boldsymbol{N}}; \widehat{\boldsymbol{M}}, \widehat{\boldsymbol{O}}\right\} = \xi^2 - \frac{2\xi^2\{\widehat{\boldsymbol{N}}^*, \widehat{\boldsymbol{N}}; \widehat{\boldsymbol{M}}, \widehat{\boldsymbol{C}}\}}{\left(1-\xi^2\right)(1+\sqrt{1+\frac{4\xi^2\{\widehat{\boldsymbol{N}}^*, \widehat{\boldsymbol{N}}; \widehat{\boldsymbol{M}}, \widehat{\boldsymbol{C}}\}}{\left(1-\xi^2\right)^2}}} \tag{6}$$

For detailed derivation, reader is referred to Barreto [12] and Duan [14].



(a) (b)

Fig 1. (a) The unifying model for image formation of central catadioptric cameras; (b) Full calibration of a hyperbolic/elliptical system using two parallel lines

2.1 Calibration Using One Set of Parallel Lines That Is Orthogonal to the Mirror Axis

Under central catadioptric projection, lines that are not coplanar with the mirror axis are projected to conics in the image. A set of parallel lines that is orthogonal to the mirror axis is represented by two line projections in Figure 1(b) which provide minimal information for the calibration task. The algorithm starts by fitting two conic curves to projected points in the image. The automatic conic fitting technique is described in section 2.3, here we assume the conic curves $\widehat{\boldsymbol{\Omega}}_{\mathbf{1}}$ and $\widehat{\boldsymbol{\Omega}}_{\mathbf{2}}$ are obtained. These two conic curves define a set of catadioptric line images which all of them intersect at two points $\widehat{\boldsymbol{F}}$ and $\widehat{\boldsymbol{B}}$. From proposition derived in [14], we know that if matrix $\boldsymbol{H}_{\boldsymbol{C}}$ is affine, the principal point $\widehat{\boldsymbol{O}}$ lies in the middle of the line segment $\overrightarrow{\widehat{\boldsymbol{F}}\widehat{\boldsymbol{B}}}$. Table 1 summarise the proposed calibration procedures.

Table 1. Algorithm of hyperbolic system calibration using the projection of at least two parallel lines orthogonal to mirror axis

Objective

Given a single image taken by hyperbolic camera, calibrate the camera and estimate the mirror parameter using line projections. The lines in the scene are assumed to be orthogonal to the mirror axis.

Algorithm

(i) Obtain at least two catadioptric line images $\widehat{\boldsymbol{\Omega}}_{\mathbf{1}}$ and $\widehat{\boldsymbol{\Omega}}_{\mathbf{2}}$ (which are the projections of lines orthogonal to the mirror axis) using conic fitting

(ii) Compute their common intersection points $\widehat{\boldsymbol{F}}$ and $\widehat{\boldsymbol{B}}$

(iii) Locate image centre $\widehat{\boldsymbol{O}}$ since $\widehat{\boldsymbol{O}}$ is collinear with $\widehat{\boldsymbol{F}}$, $\widehat{\boldsymbol{B}}$ and $|\widehat{\boldsymbol{F}}\widehat{\boldsymbol{O}}| = |\widehat{\boldsymbol{O}}\widehat{\boldsymbol{B}}|$

(iv) Obtain polar lines $\widehat{\boldsymbol{\pi}}_{\mathbf{1}}$ and $\widehat{\boldsymbol{\pi}}_{\mathbf{2}}$ of image centre $\widehat{\boldsymbol{O}}$ with respect to conic $\widehat{\boldsymbol{\Omega}}_{\mathbf{1}}$ and $\widehat{\boldsymbol{\Omega}}_{\mathbf{2}}$ using $\widehat{\boldsymbol{\pi}}_{\mathbf{1}} = \widehat{\boldsymbol{\Omega}}_{\mathbf{1}}\widehat{\boldsymbol{O}}$ and $\widehat{\boldsymbol{\pi}}_{\mathbf{2}} = \widehat{\boldsymbol{\Omega}}_{\mathbf{2}}\widehat{\boldsymbol{O}}$

(v) Compute the intersection points $\widehat{\boldsymbol{I}}_{\mathbf{1}}, \widehat{\boldsymbol{J}}_{\mathbf{1}}, \widehat{\boldsymbol{I}}_{\mathbf{2}}, \widehat{\boldsymbol{J}}_{\mathbf{2}}$ of $\widehat{\boldsymbol{\pi}}_{\mathbf{1}}$, $\widehat{\boldsymbol{\pi}}_{\mathbf{2}}$ with $\widehat{\boldsymbol{\Omega}}_{\mathbf{1}}$, $\widehat{\boldsymbol{\Omega}}_{\mathbf{2}}$, respectively

(vi) Define the absolute conic $\widehat{\boldsymbol{\Omega}}_{\infty}$ using $\widehat{\boldsymbol{I}}_{\mathbf{1}}, \widehat{\boldsymbol{J}}_{\mathbf{1}}, \widehat{\boldsymbol{I}}_{\mathbf{2}}, \widehat{\boldsymbol{J}}_{\mathbf{2}}$ and $\widehat{\boldsymbol{\pi}}_{\infty} = \widehat{\boldsymbol{\Omega}}_{\infty}\widehat{\boldsymbol{O}}$

(vii) $\boldsymbol{H}_{\boldsymbol{C}}$ is estimated from the Cholesky decomposition of $\widehat{\boldsymbol{\Omega}}_{\infty}$ since $\widehat{\boldsymbol{\Omega}}_{\infty} = \boldsymbol{H}_{\boldsymbol{C}}^{-\boldsymbol{T}}\boldsymbol{H}_{\boldsymbol{C}}^{-\mathbf{1}}$

(viii) Line $\widehat{\boldsymbol{\mu}}$ is given by $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{F}} \wedge \widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{\pi}}_{\infty} = (0, 0, 1)^{\mathrm{T}}$, so the points $\widehat{\boldsymbol{C}}_{\boldsymbol{i}}$, $\widehat{\boldsymbol{N}}_{\mathrm{i}}$, and $\widehat{\boldsymbol{M}}$ are obtained using $\widehat{\boldsymbol{C}}_{\boldsymbol{i}} = \widehat{\boldsymbol{\Omega}}_{\mathrm{i}}^{*}.\widehat{\boldsymbol{\pi}}_{\infty}$, $\widehat{\boldsymbol{N}}_{\mathrm{i}} = \widehat{\boldsymbol{\Omega}}_{\infty}^{*}.\widehat{\boldsymbol{\pi}}_{\boldsymbol{i}}$ and $\widehat{\boldsymbol{M}} = \widehat{\boldsymbol{\pi}}_{\infty} \wedge \widehat{\boldsymbol{\mu}}$, respectively

(ix) Mirror parameter ξ is then computed using equation (6)

2.2 Improving Calibration Efficiency Using a Set of Parallel Lines Coplanar with the Mirror Axis

As mentioned previously, under the central catadioptric projection, line in the 3D world is generally mapped into a small arc of a conic and sometimes occlusion also occur. This raises the difficulty of accurately estimating the conic parameters from the image itself. Even when the image points of a conic arc is manually selected, the fitting can be unreliable if the arc is smaller than 140 degrees [12][13]. This leads to inaccurate locations of $\widehat{\boldsymbol{F}}$ and $\widehat{\boldsymbol{B}}$ in the calibration method described in Table 1, as well as image centre $\widehat{\boldsymbol{O}}$. This problem can be solved by using the degenerate case of line projection (i.e. parallel lines coplanar with mirror axis). Lines coplanar with mirror axis are mapped to straight lines and intersect at principal point $\widehat{\boldsymbol{O}}$ in the catadioptric image plane (proposition derived in [14]). Lines are easier to detect and define than conic from an image since conics correspond to points in a five dimensional projective space whereas lines correspond to points in two dimensional projective space. The accuracy of line fitting for image of lines parallel to mirror axis can be also used to check if the mirror is aligned properly with the perspective camera. So the projection of a set of parallel lines coplanar with the mirror axis can provide more accurate image centre $\hat{\boldsymbol{O}}$ location than stage (i) to (iii) described in Table 1.

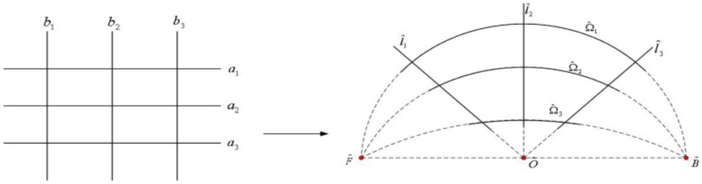


Fig. 2. Catadioptric projection of two parallel-line sets, one set is orthogonal to the mirror axis and the other set is coplanar with the mirror axis

2.3 Simultaneous Conic Detection and Grouping

In this section, since the more accurate image centre can be obtained as described above, we show how this information can be used to automate the task of conic detection and grouping.

Figure 2 illustrates the catadioptric projection of two parallel-line sets. Parallel lines $\boldsymbol{a_1}$, $\boldsymbol{a_2}$ and $\boldsymbol{a_3}$ are orthogonal to the mirror axis and parallel lines $\boldsymbol{b_1}$, $\boldsymbol{b_2}$ and $\boldsymbol{b_3}$ are planar with the mirror axis. If $\boldsymbol{a_1}$, $\boldsymbol{a_2}$ and $\boldsymbol{a_3}$ are assumed to have equal length, then their mapped image – conic curves $\hat{\boldsymbol{\Omega}}_\mathbf{1}$, $\hat{\boldsymbol{\Omega}}_\mathbf{2}$ and $\hat{\boldsymbol{\Omega}}_\mathbf{3}$ will have different arc length. $\boldsymbol{b_1}$, $\boldsymbol{b_2}$ and $\boldsymbol{b_3}$ are mapped to $\hat{\boldsymbol{l}}_\mathbf{1}$, $\hat{\boldsymbol{l}}_\mathbf{2}$ and $\hat{\boldsymbol{l}}_\mathbf{3}$ in the catadioptric image. The principal point $\hat{\boldsymbol{O}}$ can be computed by solving

$$\min_{\hat{\boldsymbol{O}}} \sum_{i=1}^{n} \left(\hat{\boldsymbol{l}}_i^T \hat{\boldsymbol{O}}\right)^2 \qquad (7)$$

Table 2 summarise the proposed conic fitting approach.

Table 2. Conic fitting algorithm for improving fitting accuracy and enabling automatic processing

Objective

Given the catadioptric image of two sets of parallel lines, one set orthogonal to the mirror axis and the other coplanar with the mirror axis, fit conics to the arc segments and estimate the common intersection $\hat{\boldsymbol{F}}$ and $\hat{\boldsymbol{B}}$.

Algorithm

(i) Compute the principal point $\hat{\boldsymbol{O}}$ using equation (7)

(ii) Fit conic $\hat{\boldsymbol{\Omega}}_\mathbf{1}$ to the arc with largest degree value

(iii) Estimate $\hat{\boldsymbol{F}}_\mathbf{1}$ and $\hat{\boldsymbol{B}}_\mathbf{1}$ using constraints: (1) $|\hat{\boldsymbol{F}}_\mathbf{1}\hat{\boldsymbol{O}}| = |\hat{\boldsymbol{O}}\hat{\boldsymbol{B}}_\mathbf{1}|$; (2) $\hat{\boldsymbol{F}}_\mathbf{1}$, $\hat{\boldsymbol{O}}$ and $\hat{\boldsymbol{B}}_\mathbf{1}$ are collinear; and (3) $\hat{\boldsymbol{F}}_\mathbf{1}$ and $\hat{\boldsymbol{B}}_\mathbf{1}$ lie on the conic fitted to $\hat{\boldsymbol{\Omega}}_\mathbf{1}$

Table 2. (*Continued*)

(iv)	The value of $\widehat{\boldsymbol{F}}_1$ and $\widehat{\boldsymbol{B}}_1$ is assigned to $\widehat{\boldsymbol{F}}'$ and $\widehat{\boldsymbol{B}}'$
(v)	Estimate the next conic $\widehat{\boldsymbol{\Omega}}_i$ using $\widehat{\boldsymbol{F}}'$, $\widehat{\boldsymbol{B}}'$ and image points of available arc, $\widehat{\boldsymbol{F}}'$, $\widehat{\boldsymbol{B}}'$ can be weighted according to the degree span of the first arc
(vi)	Re-estimate all fitted conics with the updated $\widehat{\boldsymbol{F}}'$, $\widehat{\boldsymbol{B}}'$
(vii)	Compute $\widehat{\boldsymbol{F}}_i$ and $\widehat{\boldsymbol{B}}_i$ that lie on $\widehat{\boldsymbol{\Omega}}_i$ exactly using constraints from (iii) for each conic fitted so far
(viii)	Update $\widehat{\boldsymbol{F}}'$ and $\widehat{\boldsymbol{B}}'$ with the means of all $\widehat{\boldsymbol{F}}_i$ and $\widehat{\boldsymbol{B}}_i$ and go back to stage (v)
(ix)	Stop the iteration when $\widehat{\boldsymbol{F}}'_{i+1} - \widehat{\boldsymbol{F}}_i'$ or $\widehat{\boldsymbol{B}}'_{i+1} - \widehat{\boldsymbol{B}}_i'$ is smaller than τ, or all arc segments has been fitted

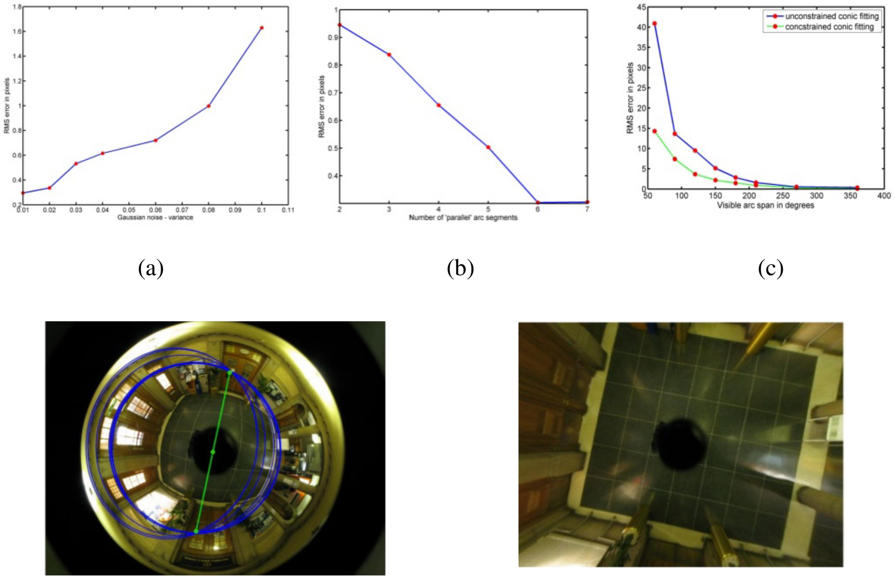
3 Experiment Results

A number of experiments are carried out on both simulated and real data. The proposed method is evaluated with respect to noise sensitivity, occlusion and the number of detected segments. Since the location of intersecting points $\widehat{\boldsymbol{F}}$, $\widehat{\boldsymbol{B}}$ and principal point (O_x, O_y) are the keys to unlock the whole calibration of catadioptric system as described in Table 1, the estimated point coordinates are compared to its ground truth value. The difference between constrained conic fitting described in Table 2 and free conic fitting (i.e. simply fit a conic to each arc) is also tested.

3.1 Calibration with Simulated Data

The calibration parameters of the simulated catadioptric camera are set as: aspect ratio r = 1, skew factor s = 0, effective focal length f_e = 450, image centre $O_x = 400$ and $O_y = 300$, image size 800×600 and mirror parameter $\xi = 0.9886$. First of all, a set of N = 5 lines are generated by choosing five normal vectors $\boldsymbol{n}$ from the unit sphere. To make sure this set of lines are parallel and orthogonal to the mirror axis, the unit sphere is expressed spherical coordinate system (θ, φ, r) where $\theta \in [0, 2\pi]$, $\varphi \in [0, \pi]$ and r = 1, then the value of φ of the normal vector are randomly generated while keeping the value of θ unchanged.

The first experiment is to test the noise sensitivity of proposed method. Gaussian noise with zero mean and σ standard deviation is added to the synthetic image. The noisy simulated images with value of variance varying from 0.01 to 0.1. For each noise level, results are compared to the ground truth and RMS error are computed from 200 independent trials. In Figure 3(a), the algorithm's robustness against noise is evaluated. The second experiment conducted is to test how the calibration accuracy change against the number of arc segments available. Here, the value of Gaussian noise variance is set to σ = 0.05. The results in Figure 3(b) shows that calibration accuracy increase with the number of arc segments detected and sub-pixel accuracy can be reached by five conics. In the third experiment, we test the robustness of the proposed method against occlusion. The visible conic segments are varied from 360 to 60 degrees. Here, our proposed constrained conic fitting are compared to conic fitting independently to each visible arc. As can be seen from the results showing in Figure 3(c), the proposed algorithm shows good improvement on calibration accuracy compared to unconstrained conic fitting, especially when occlusion is more than 180 degrees. For all the experiment described above, conic fitting technique LMS – normal least squares minimisation is chosen to be used in our experiment. This is because our major target is to show the use of two special parallel line sets in catadioptric calibration and how the conic fitting can be constrained to improve calibration accuracy.



(a) (b) (c)

(d) (e)

Fig. 3. (a)(b)(c) Experimental results from simulated data; (d) Conic detected and grouped to locate the vanishing points; (e) Rectification of floor using the estimated calibration results

3.2 Calibration with Real Data

For the experiment with real data, we used a perspective camera with a hyperbolic mirror. The hyperbolic mirror is commercially available named The 0-360 Panoramic Optic, which has the vertical field of view (FOV) of 115 degrees. Figure 3(d) shows the detected conics corresponding to a set of parallel lines orthogonal to the mirror axis. Figure 3(e) shows the rectification using the calibration results. During the experiment, it is also found that the absolute conic $\widehat{\boldsymbol{\Omega}}_\infty$ has to be positive definite to enable calibration for the method described in Table 1. However, when noise is large, the algorithm is then unstable. Luckily, as long as the intersection points of conics for parallel line set are detected, the problem can be avoid using similar method as proposed by Geyer and Daniilidis [11].

4 Summary

This paper first derived that catadioptric camera can be calibrated using the projected image of only one set of parallel lines. This set of lines are assumed to be orthogonal to the mirror axis. To enable automatic calibration from a single catadioptric image, automatic detection and grouping of conics corresponding to the parallel lines are required. This can be achieved using another set of parallel lines which are coplanar with the mirror axis. The projection of these lines all intersect at the principal point. The experiment results show that larger number of detected 'parallel' conics can improve the accuracy of vanishing points location. The proposed constraint on conic fitting can perform twice as good as unconstraint conic fitting especially when arc occlusion is large. The research work described in this paper is only the first step of exploiting the use of featured parallel line under catadioptric projection. Its usage can be extended further in applications such as 3d reconstruction and extracting metric information from the catadioptric image.

References

1. Barreto, J., Araujo, H.: Direct Least Square Fitting of Paracatadioptric Line Images. In: Computer Vision and Pattern Recognition Workshop, pp. 78–87 (2003)
2. Kang, S.B.: Catadioptric Self-calibration. In: IEEE Conference on CVPR, pp. 201–207 (2000)
3. Strelow, D., Mishler, J., Koes, D., Singh, S.: Precise Omnidirectional Camera Calibration. In: IEEE Conference on CVPR, pp. 689–694 (2001)
4. Scaramuzza, D., Martinelli, A., Siegwart, R.: A Flexible Technique for Accurate Omnidirectional Camera Calibration and Structure from Motion. In: IEEE Conference on Computer Vision Systems (2006)
5. Gasparini, S., Sturm, P., Barreto, J.P.: Plane-based Calibration of Central Catadioptric Cameras. In: IEEE Conference on Computer Vision, pp. 1195–1202 (2009)
6. Mei, C., Rives, P.: Single View Point Omnidirectional Camera Calibration from Planar Grids. In: IEEE Conference on Robotics and Automation, pp. 3945–3950 (2007)
7. Micusik, B., Pajdla, T.: Para-catadioptric Camera Auto-calibration from Epipolar geometry. In: ACCV (2004)
8. Ying, X., Hu, Z.: Catadioptric Camera Calibration Using Geometric Invariants. IEEE Transactions on Pattern Analysis and Machine Intelligence 26, 1260–1271 (2004)
9. Ying, X., Zha, H.: Identical Projective Geometric Properties of Central Catadioptric Line Images and Sphere Images with Applications to Calibration. International Journal of Computer Vision 78, 89–105 (2008)
10. Ying, X., Zha, H.: Simultaneously Calibrating Catadioptric Camera and Detecting Line Features Using Hough Transform. In: IEEE Conference on Intelligent Robots and Systems, pp. 1343–1348 (2005)
11. Geyer, C., Daniilidis, K.: Catadioptric Camera Calibration. In: IEEE Conference on Computer Vision, pp. 398–404 (1999)
12. Barreto, J.: General Central Projection Systems, Modeling, Calibration and Visual Servoing. PhD thesis, Dept. of Electrical and computer engineering, University of Coimbra (2003)
13. Zhang, Z.: Parameter Estimation Techniques: A tutorial with application to conic fitting. Rapport de Recherche No. 2676, INRIA (1995)
14. Duan, W.: Vanishing Point Detection and Camera Calibration. PhD thesis, Dept. of Electrical and Electronic Engineering, University of Sheffield (2011)

Color Histogram-Based Image Segmentation

Giuliana Ramella and Gabriella Sanniti di Baja

Istituto di Cibernetica "E.Caianiello", CNR
Via Campi Flegrei 34, 80078 Pozzuoli, Naples, Italy
{g.ramella,g.sannitidibaja}@cib.na.cnr.it

Abstract. An algorithm is presented to segment a color image based on the 3D histogram of colors. The peaks in the histogram, i.e., the connected components of colors with locally maximal occurrence, are detected. Each peak is associated a representative color, which is the color of the centroid of the peak. Peaks are processed in decreasing occurrence order, starting from the peak with the maximal occurrence, with the purpose of maintaining only the representative colors corresponding to the dominant peaks. To this aim, each analyzed peak groups under its representative color those colors, present in the histogram and that have not been grouped to any already analyzed peak, such that their distance from the centroid of the peak is smaller than a priori fixed value. At the end of the grouping process, a number of representative colors, generally substantially smaller than the number of initial peaks, is obtained, which are used to identify the regions into which the color image is segmented. Since the histogram does not take into account spatial information, the image is likely to result over-segmented and a merging step, based on the size of the segmentation regions, is performed to reduce this drawback.

1 Introduction

Image segmentation is a key process in pattern recognition and computer vision, since the quality of any image analysis and understanding task is highly conditioned by the quality of the segmentation results. Purpose of segmentation is to partition an input image into a number of disjoint regions, each of which should ideally correspond to one of the regions that a human observer perceives in the scene. The regions of the partition should be such that pixels in a given region are similar as regards a given property, e.g., color, intensity, or texture, while pixels belonging to adjacent regions should differ from each other significantly as regards that property.

Different segmentation approaches have been suggested in the literature, such as histogram thresholding, feature clustering, multiresoltion representation, region-based methods, fuzzy techniques, neural networks (see, e.g., [1-9]).

The majority of the published papers deal with gray level images or suggest color image segmentation techniques that are based on gray level image segmentation schemes [10]. The three components of the image in the selected color space are processed individually as gray level images, or by considering pair-wise color projections [11] and the results are then combined to originate the segmented color image.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 76–83, 2011.

One of the main reasons for resorting to gray level segmentation of the individual color components is the noticeably higher computational complexity of the high-dimensional color space. However, since the human visual system is able to distinguish many more color shades than shades of gray, in some cases color is essential to correctly identify all objects in a scene. Thus, by taking into account that color information as a whole is likely to be lost when the color is projected onto three components and also that an increasing number of processing hardware able to deal with the computational complexity due to the high-dimensional color space are available, methods processing directly color images are of interest.

Color segmentation can be seen as a clustering problem in the 3D space, where the coordinate axes are the color components and each point represents one of the colors in the image. Clustering is accomplished by assigning a set of observations to well separated clusters, where the observations in a given cluster are similar in terms of certain features. Image features are generally based on color or texture and are computed within a small size window centered on each pixel to be classified. The most commonly used clustering methods are the K-means [12] and the fuzzy C-means [13]. Particularly the fuzzy C-means algorithm has been widely employed due to its ability to produce rather compact regions in the segmented image and thanks to the simplicity of implementation. However, the method requires an initial decision on the number of clusters and an appropriate distribution of the initial cluster centroids, which may affect the quality of the resulting segmentation. Some methods aiming at the solution of the above problems have been recently published, e.g. [14,15], but the proposed suggestions do not always efficiently overcome the problems.

In this paper, we suggest a color image segmentation scheme based on the use of the 3D histogram of the image. The underlying assumption is that each region of the image characterized by almost uniform color corresponds in the histogram to a group of bins including a dominant peak.

To identify the dominant peaks, we process all peaks found in the histogram in decreasing occurrence order, starting from the peaks with maximal occurrence. Each peak gathers the colors whose distance from the peak is smaller than an a priori fixed value, which is set based on color distribution. All gathered colors are associated the same representative color. Segmentation is obtained by changing the true color of each pixel of the input image with the appropriate representative color. Due to the fact that the histogram does not take into account spatial information, the image resulting after the true colors have been replaced by the representative colors is likely to be over-segmented. Thus, a merging step, based on the size of the segmentation regions, is then taken into account, aimed at reducing over-segmentation.

With respect to C-means, we do not need to fix a priori the number of dominant peaks. Moreover, we do not require any particular initial distribution of the centroids. A positive aspect of our method is that color information is employed without resorting to the separation of the three color components or to the use of projections onto suitable color planes. Though working with the 3D histogram, the method is not highly expensive from the computational point of view and produces in the average satisfactory results.

The paper is organized as follows. Some basic notions are given in Section 2. The algorithm is described in Section 3. Experimental results are shown in Section 4. Finally, concluding remarks are given in Section 5.

2 Preliminaries

We work with RGB images and interpret colors as three-dimensional vectors, with each vector element having an 8-bit dynamic range.

The 3D histogram H of a 2D color image I is computed by representing the RGB color space as a three-dimensional cube in a cubic grid with values ranging along each axis from 0 to 255. The voxels of H are initially set to 0. For each pixel of I with color (x, y, z), the voxel in position (x, y, z) of H has its value increased by 1. When all pixels of I have been inspected, each voxel of the three-dimensional cube H counts the number of pixels of I with the same color.

The 3×3 (3×3×3) neighborhood of a pixel (voxel) p of I (H) is the set including the 8 (26) neighbors of p. In the following, we will use for short the same symbol p to refer to both a pixel of I (voxel of H) and to its color (occurrence).

To identify the peaks of H, we find the connected components of voxels with locally maximal occurrence, i.e., any connected set C of voxels, all with the same occurrence p, such that any voxel that is not in C but has at least a neighbor in C has occurrence smaller than p. A peak may consist of a single voxel or of a connected component of voxels with the same occurrence in H. The latter case occurs when similar colors, corresponding to neighboring voxels in H, are equally frequent. Synthetical images can be built where the furthest voxels of a peak correspond to very different colors. However, for natural images this does not generally happens and each peak of H includes only voxels corresponding to colors rather similar to each other.

The representative color associated with a peak consisting of a single voxel in position (x, y, z) is obviously the color characterized by R=x, G=y and B=z. In turn, for a peak consisting of a connected component of voxels, the representative color associated to the peak is the color of the centroid of the peak.

3 The Algorithm

Our segmentation algorithm consists of two steps. The first step identifies the main colors present in the image and builds segmentation regions where pixels in each region are closer to one of these colors than to any other color. The second step performs merging, based on the size of segmentation regions, to reduce the number of regions in the resulting segmented image.

During the first step, the histogram H is built and all the peaks are identified. Peaks of H with occurrence smaller than a parameter θ, whose value is fixed based on color distribution, are not analyzed. Hence, these peaks for sure will not be regarded as dominant peaks, even if they are grouped into large connected components including many voxels with the same locally maximal occurrence. This is done to avoid the detection of a very large number of scarcely significant regions in the resulting segmented image and to decrease the computation time. Of course, we are aware that disregarding a number of peaks, even if characterized by limited occurrence, may result at the end of the first step in an image where not all pixels of I are assigned to a segmentation region. Pixels of I that cannot be assigned a representative color and, hence, are not assigned to any segmentation region, are treated during the second step of the process.

The value of θ should be selected by taking into account the occurrence of all peaks in the histogram. In our opinion, the value of θ should be a small percentage of the mean value of peak occurrence. Let μ indicate the arithmetic mean of the occurrences of the peaks of H, then we have experimentally found that θ=1% μ as default value for the parameter θ produces in the average satisfactory results.

A second parameter τ is used to group around any peak the voxels of H whose colors do not dramatically differ from the color of the peak. This is done by associating the same representative color characterizing any peak to all voxels of H that are grouped with that peak. The value of τ depends on the maximal dissimilarity that the user accepts for colors to be grouped together. By taking into account the maximal Euclidean distance between different colors in the 3D histogram and the minimum number of colors expected to characterize a complex color image, we suggest τ=50 as default value. Also this default value has been determined experimentally.

The peaks are examined in decreasing occurrence order, since the relevance of a peak is strongly conditioned by the number of times the color associated with the peak appears in the image I. In principle, we regard as dominant all the detected peaks. However, peaks whose associated representative colors do not strongly differ from each other may result to be merged into a unique group. Thus the final number of groups, and hence of representative colors, is likely to be smaller than the number of detected peaks. The larger is τ, the higher is the possibility to merge peaks. All voxels of H that have value different from zero and a distance from the centroid of the peak smaller than τ are associated to the current peak. Voxels of H already associated to a given peak are not taken into account when other peaks are examined.

Obviously, large values of τ may cause grouping of colors that a user would perceive as dissimilar. On the other hand, small values of τ not only produce over-segmented resulting images where a huge number of colors is used, but also risk to leave large portion of the input image I not assigned to any region. This would be the case for the pixels of I whose colors are at distance larger than τ from any peak detected in H. Thus, selection of θ and τ is a key point to obtain a good result. For the time being, we have experimentally found that the best values of θ and τ for each image, obtained by running the algorithm with different values for the parameters and by comparing the correspondingly obtained results, coincide in the average with the suggested default values.

Once all peaks have been examined, the image I is inspected. Each pixel of I with color (x, y, z) is set to the representative color associated with the voxel in position (x, y, z) of H. Pixels of I for which no representative color is found in H are set to a special value in I, used to point out that these pixels remain temporarily as unassigned. We point out that differently from the representative colors, the special value may be assigned to pixels whose colors differ from each other significantly more than τ. Thus, it is very important to select the values for θ and τ in such a way to be sure that only a few pixels, possibly sparse or grouped in small size regions, remain unassigned.

Clearly, the same representative color, or the special value used for the unassigned pixels, may be set in correspondence with pixels not necessarily belonging to the same connected region of I. In fact, the histogram only counts the number of times that a color appears in I but does not take into account any spatial information. Thus, also the grouping process that gathers colors around a peak does not take into account spatial information.

In Fig. 1 top the image "church" used as running example is shown, while the image resulting at the end of the first step of the segmentation algorithm is shown in Fig. 1 bottom left. The input image includes 23326 different colors. Out of the 1487 peaks detected in the 3D histogram characterized by μ=300, only 13 final representative colors are obtained by using the default values for θ and τ (namely θ=3 and τ=50).

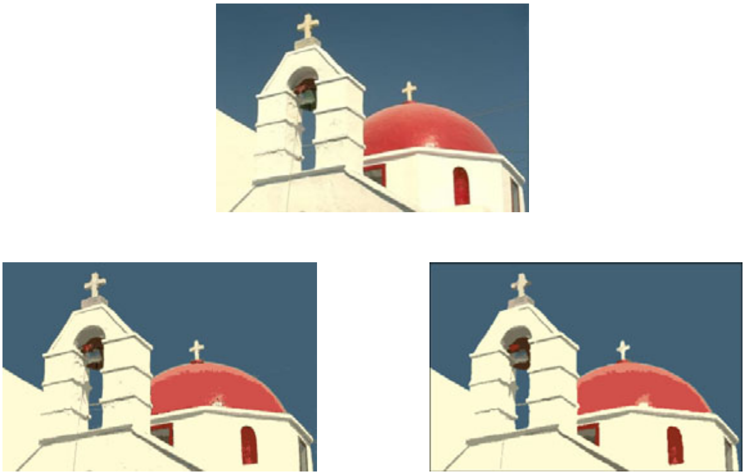


Fig. 1. The input image "church" (top), the image resulting after the first step of segmentation (bottom left) and the final segmented image (bottom right)

The second step of the segmentation algorithm is aimed at region merging. Merging is done both to take care of unassigned pixels and to reduce over-segmentation.

Preliminarily, connected component labeling is accomplished to distinguish all regions of the partition of I. For each connected component, the area of the region is recorded.

During one inspection of I, pixels belonging to a connected component having area smaller than an a priori fixed value γ are set to zero. These pixels are successively assigned to the adjacent region with which they have the largest number of neighbors. Based on the experiments that we have carried on, we suggest as default value for γ 25% of the arithmetic mean A of the area of the segmentation regions.

Merging may cause a reduction in the number of representative colors. This happens whenever pixels of I that are assigned a given representative color are all grouped into connected components characterized by area smaller than γ. Of course, the number of segmentation regions is generally larger than the number of representative colors.

For the running example, the resulting segmented image is shown in Fig. 1 bottom right. The first step of the process generated 2644 connected components with A=60, which were reduced to only 96 final segmentation regions, after the merging step accomplished by using the default value γ=15.

4 Experimental Results

We have applied our segmentation algorithm to a collection of images with different size and color distribution, taken from available databases, e.g., [16-20]. A small dataset including four images, used as test images together with the image "church" to show the performance of our method, is given in Fig. 2.

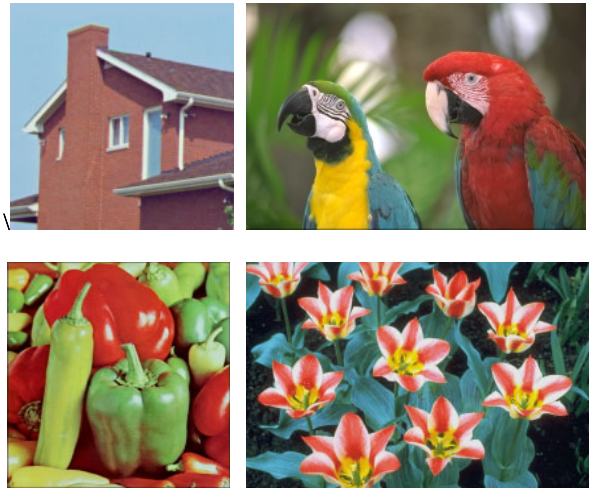


Fig. 2. Images "house", "parrots", "peppers" and "tulips"

The resulting segmentations for the images "house", "parrots", "peppers" and "tulips" are shown in Fig.3. For each test image, the number of input colors, the number of peaks detected in H, the number of dominant peaks, the number of final representative colors, the number of partition regions before merging and the final number of partition regions are given in Table 1.

The five images in Table 1 differ significantly for the number of colors, ranging from only 256 colors for "tulips" to 111344 colors for "peppers". Also the number of peaks initially detected in each histogram significantly differs, while the number of dominant peaks and of final representative colors does not largely differ for the different images. We also point out the effect of merging by comparing the number of regions found before and after merging.

Table 1. Results for the test images

image	input colors	peaks	dominant peaks	final repr. colors	regions before merging	final regions
church	23326	1487	14	13	2644	96
house	33847	9633	90	32	4511	265
parrots	49942	48685	74	57	10710	1079
peppers	111344	15884	57	36	11854	510
tulips	256	256	23	22	12254	551

The experiments show that the salient regions of images are effectively extracted and that the segmentation results are close to those expected by taking into account human perception. As already pointed out, the number of segmentation regions is generally significantly larger than the number of final representative colors. This is particularly the case for images like "parrots", where colors of the original image that have been grouped into the same representative color are spread all over the image.

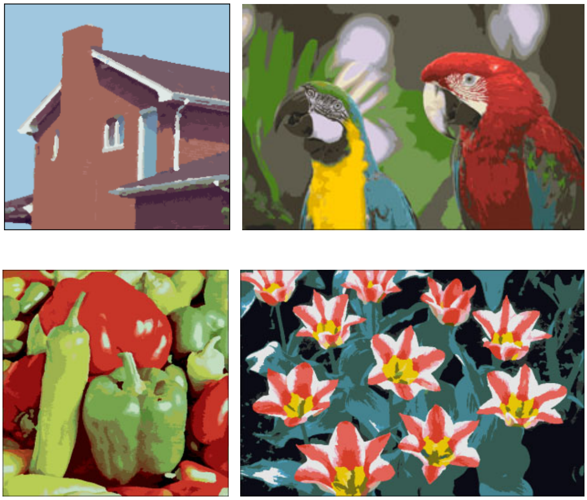


Fig. 3. Segmentations for "house", "parrots", "peppers" and "tulips

5 Concluding Remarks

We have presented an algorithm for image segmentation based on the 3D histogram of colors. The peaks of the histogram are detected and are processed in decreasing occurrence order starting from the peak with maximal occurrence. A peak groups under the same representative color all colors with distance from the peak smaller than an a priori fixed value, set depending on color distribution. At the end of this process, only dominant peaks survive, which are generally less numerous than the initially detected peaks. The resulting representative colors are used to identify the regions into which the input image is segmented. Since the histogram does not include spatial information, the image is likely to result over-segmented. Thus, a merging step is done to reduce over-segmentation.

To reduce the computation time involved by the use of the 3D histogram, peaks with small occurrence are not processed. This may cause a number of generally sparse pixels of the input image to be not assigned to any segmentation region. The merging step, aimed at merging regions with small area to the adjacent larger regions, also allows us to treat unassigned pixels.

References

1. Tan, K.S., Isa, N.A.M.: Color image segmentation using histogram thresholding – Fuzzy C-means hybrid approach. Pattern Recognition 44, 1–15 (2011)
2. Aghbari, Z.A., Al-Haj, R.: Hill-manipulation: An effective algorithm for color image segmentation. Image and Vision Computing 24, 894–903 (2006)
3. Huang, R., Sang, N., Luo, D., Tang, Q.: Image segmentation via coherent clustering in L*a*b* color space. Pattern Recognition Letters 32, 891–902 (2011)
4. Kuan, Y.H., Kuo, C.M., Yang, N.C.: Color-Based Image Salient Region Segmentation Using Novel Region Merging Strategy. IEEE Transactions on Multimedia 10(5) (2008)
5. Jung, C.R.: Unsupervised multiscale segmentation of color images. Pattern Recognition Letters 28, 523–533 (2007)
6. Celik, T., Tjahjadi, T.: Unsupervised colour image segmentation using dual-tree complex wavelet transform. Computer Vision and Image Understanding 114, 813–826 (2010)
7. Wangenheim, A.V., Bertoldi, R.F., Abdala, D.D., Sobieranski, A., Coser, L., Jiang, X., Richter, M.M., Priese, L., Schmitt, F.: Color image segmentation using an enhanced Gradient Network Method. Pattern Recognition Letters 30, 1404–1412 (2009)
8. Sowmya, B., Rani, B.S.: Colour image segmentation using fuzzy clustering techniques and competitive neural network. Applied Soft Computing 11, 3170–3178 (2011)
9. Araújo, A.R.F., Costa, D.C.: Local adaptive receptive field self-organizing map for image color segmentation. Image and Vision Computing 27, 1229–1239 (2009)
10. Cheng, H.D., Jiang, X.H., Sun, Y., Wang, J.: Color image segmentation: advances and prospects. Pattern Recognition 34, 2259–2281 (2001)
11. Lézoray, O., Charrier, C.: Color image segmentation using morphological clustering and fusion with automatic scale selection. Pattern Recognition Letters 30, 397–406 (2009)
12. Lloyd, S.P.: Least squares quantization in PCM. IEEE Trans. Inf. Theory 28(2), 129–136 (1982)
13. Berkhin, P.: Survey of clustering data mining techniques. Accrue Software (2002)
14. Chen, T.W., Chen, Y.L., Chien, S.Y.: Fast Image Segmentation Based on K-Means Clustering with Histograms in HSV Color Space. In: Proc. of IEEE 10th Workshop on Multimedia Signal Processing, pp. 322–325 (2008)
15. Yu, Z., Au, O.C., Zou, R., Yu, W., Tian, J.: An adaptive unsupervised approach toward pixel clustering and color image segmentation. Pattern Recognition 43, 1889–1906 (2010)
16. http://www.hlevkin.com/TestImages/
17. http://sipi.usc.edu/database/
18. http://r0k.us/graphics/kodak/
19. http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/
20. http://decsai.ugr.es/cvg/dbimagenes/

Arc Segmentation in Linear Time*

Thanh Phuong Nguyen and Isabelle Debled-Rennesson

ADAGIo team, LORIA, Nancy University
54506 Vandoeuvre-lès-Nancy, France
{nguyentp,debled}@loria.fr

Abstract. A linear algorithm based on a discrete geometry approach is proposed for the detection of digital arcs and digital circles using a new representation of them. It is introduced by inspiring from the work of Latecki [1]. By utilizing this representation, we transform the problem of digital arc detection into a problem of digital straight line recognition. We then develop a linear method for arc segmentation of digital curves.

1 Introduction

The digital arcs and circles are basic geometric objects of which the recognition is an interesting topic. In the literature, some methods have been proposed for the recognition of digital circles. Nakamura et al. [2] proposed a recursive algorithm for determining the center of a digital circle, but its complexity is exponential in the general case. Kim [3,4] proposed several results on digital disks. The first result [3] detects if a set of grid points in a $N \times N$ image is a digital disk with complexity $O(n^3)$. The second result [4] reduces this task to $O(n^2)$. Based on the classical separating arc problem, Kovalevsky [5] (resp. Fisk [6]) proposed an algorithm for the recognition of a digital disk in $O(n^2 \log n)$ (resp. $O(n^2)$) time. Coeurjolly [7] transformed the problem of circle recognition into a problem of search a 2D point that belongs to the intersection of n^2 half-plane. Sauer [8] (resp. Damaschke [9]) presented a linear algorithm to decide if a curve is a digital circle (resp. arc) based on Megiddo's algorithm [10]. Worring [11] introduced a digital circle segmentation method by using a fixed size window process. Roussillon [12] proposed a linear algorithm of circle recognition in 3 particular cases.

We present in this paper a linear method for the detection of digital circles or digital arcs based on a discrete geometry approach. Firstly, a polygonalization is applied in linear time on the input curve [13]. Secondly, we use a transform proposed by Latecki et al. [1] to represent the obtained polygon in a novel space called ***tangent space***. We show that a sequence of chords of a circle will correspond to a sequence of collinear points in the tangent space. So the problem of arc/circle detection can be considered as a problem of digital straight line recognition.

This paper is organized as follows. Section 2 recalls some definitions concerning digital circles and blurred segments. The next section presents a technique

[*] We would like to thank Eric Domenjoud for his valuable comments.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 84–92, 2011.

to transform an arc into the tangent space and proposes some principal properties of the arc in this representation. Section 4 proposes a linear algorithm for the detection of digital arcs or digital circles. In Section 5, we present a linear method for the segmentation of a curve into arcs and some experimentations.

2 Discrete Circle and Blurred Segment

Discrete circle: In the literature, there exist several definitions of discrete circle. They are proposed by considering a real circle on the grid digitization. The difference among them is the process of discretization. Nakamura et al. [2] considered a discretization of a real circle by the points of $\mathbb{Z}^2$ that are the nearest points of that circle. Kim [3] proposed a definition of discrete circle as a boundary of a digital disk superimposed by a real circle. Andres [14] used an arithmetic approach to define a digital circle as a sequence of points superimposed by a ring.

Discrete line and blurred segment: The notion of blurred segment [13] was introduced from the notion of arithmetical discrete line. An ***arithmetical discrete line***, noted $D(a, b, \mu, \omega)$, is the set of points $(x, y) \in \mathbb{Z}^2$ that verifies: $\mu \leq ax - by < \mu + \omega$ with a main vector (b, a), lower bound μ and thickness ω . A ***width*** ν ***blurred segment*** (BS) is a set of points $(x, y) \in \mathbb{R}^2$ that is optimally bounded (see [13] for more details) by a discrete line $D(a, b, \mu, \omega)$ verifying $\frac{\omega-1}{max(|a|,|b|)} \leq \nu$. Fig. 1 shows a BS of with 1.25 (the sequence of gray points) whose optimal bounding line is $\mathcal{D}(5, 8, -8, 11)$. A linear method for recognition of BS has been also proposed in [13].

3 Arc Representation in Tangent Space

Modified tangent space representation: We recall in this section some notions concerning a representation of a polygon in the tangent space. Latecki et al. [1] proposed ***the tangent space representation*** as a tool of similarity measure for shape matching. Inspired from this representation, we propose a modified tangent space to represent a polygonal curve. The difference is that we do not normalize the axis $0x$ in the tangent space. Let $C = \{C_i\}_{i=0}^{n}$ be a polygonal curve, $\alpha_i = \angle(\overrightarrow{C_{i-1}C_i}, \overrightarrow{C_iC_{i+1}})$ and l_i - the length of the line segment $C_iC_{i+1}, i \in \{0, \ldots, n-1\}$. If C_{i+1} is on the right of $\overrightarrow{C_{i-1}C_i}$ then $\alpha_i > 0$, otherwise $\alpha_i < 0$. From now, we denote $P.x$ (resp. $P.y$) to indicate the x (resp. y)-coordinate of point P. We consider a transformation that associates the polygon C of $\mathbb{Z}^2$ to a polygon of $\mathbb{R}^2$ that is constituted by line segments $T_{i2}T_{(i+1)1}, T_{(i+1)1}T_{(i+1)2}$ for i from 0 to $n-1$ with

$$\begin{aligned}
&T_{02} = (0, 0),\\
&T_{i1} = (T_{(i-1)2}.x + l_{i-1}, T_{(i-1)2}.y),\ i \text{ from } 1 \text{ to } n,\\
&T_{i2} = (T_{i1}.x, T_{i1}.y + \alpha_i),\ i \text{ from } 1 \text{ to } n-1.
\end{aligned}$$

Properties of arcs in the modified tangent space representation. The theorem below allows us to study the properties of a representation in the tangent space of a polygon that corresponds to an arc or a circle.

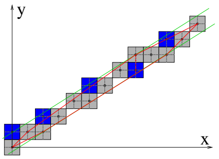


Fig. 1. A BS [13]

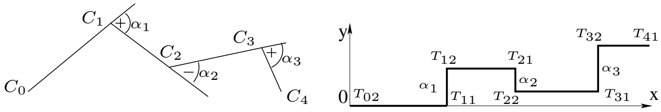


Fig. 2. Transformation to the tangent space: on the left, the input polygonal curve and on the right, its tangent space representation

Theorem 1. *Let $C = \{C_i\}_{i=0}^{n}$ be a polygon, $\alpha_i = \angle(\overrightarrow{C_{i-1}C_i}, \overrightarrow{C_iC_{i+1}})$. The length of C_iC_{i+1} is l_i, for $i \in \{0, \ldots, n-1\}$. The vertices of C are on a real arc of radius R and of center O such that $\angle C_iOC_{i+1} \leq \frac{\pi}{4}$ for $i \in \{1, \ldots, n-1\}$. This results below is obtained.*

$$\frac{1}{R} < \frac{\alpha_i}{\frac{l_i + l_{i+1}}{2}} < \frac{1}{0.9742979R}$$

Proof. Let us consider figure 3. We have $\alpha_i = \angle C_iOH_{i-1} + \angle C_iOH_i$. We denote that $\alpha_{i1} = \angle C_iOH_{i-1}$ and $\alpha_{i2} = \angle C_iOH_i$. Moreover, $\angle C_1OH_0 = \frac{\angle C_0OC_1}{2} \leq \frac{\pi}{8}$, $\angle C_1OH_1 = \frac{\angle C_1OC_2}{2} \leq \frac{\pi}{8}$. In addition, we have $\sin \angle C_1OH_0 = \frac{l_0}{2R}$, $\sin \angle C_1OH_1 = \frac{l_1}{2R}$. Therefore, $\frac{l_0 + l_1}{2R} = \sin \alpha_{11} + \sin \alpha_{12}$. Similarly, we have $\frac{l_{i-1} + l_i}{2R} = \sin \alpha_{i1} + \sin \alpha_{i2}$, for $i \in \{1, \ldots, n-1\}$. Because of $x \geq \sin x \geq x - \frac{x^3}{6}$ with $x > 0$, we have $\alpha_{i1} > \sin \alpha_{i1} > \alpha_{i1}(1 - \frac{\alpha_{i1}^2}{6}) > \alpha_{i1}(1 - \frac{\frac{\pi}{8}^2}{6}) > 0.9742979\alpha_{i1}$. Similarly $\alpha_{i2} > \sin \alpha_{i2} > 0.9742979\alpha_{i2}$. Therefore, we have $\alpha_i > \frac{l_{i-1} + l_i}{2R} > 0.9742979\alpha_i$. This theorem is proved.

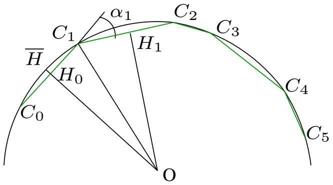


Fig. 3. Property of a set of sequential chords of a partial circle

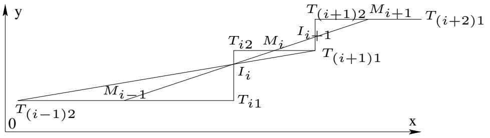


Fig. 4. Property of a polygon in our modified tangent space representation

This theorem allows to deduce that the corresponding curve of midpoints of $T_{(i-1)2}T_{i1}$, $1 \leq i \leq n$ in the tangent space of the curve C is quasi collinear. From now on, the midpoint curve is called MpC. In addition, the more $\sin \alpha_i$ closes to α_i, $1 \leq i < n$, the more MpC is collinear. Therefore, we can decide if a digital curve approximates an arc of circle by verifying the collinearity of its MpC in the tangent space. A qualitative study on this approximation will be also considered in Section 5.

4 Arc Segmentation

Detection of digital arcs. Thanks to Theorem 1, we introduce now an heuristic algorithm (see algo. 1) for deciding if a digital curve is an arc. Our main idea is to work on the representation of a digital curve in the modified tangent space. In this representation, the set of midpoints $MpC = \{M_i\}_{i=0}^{n-1}$ (see the above section) will be constructed. And we will use a linear procedure [13] to test the collinearity of these points. If the response is positive, we consider that the input digital curve is a digital arc (a partial circle).

Algorithm 1. Detection of a digital arc/circle

```
Data: $P = \{P_i\}_{i=0}^{n}$ digital curve, $\alpha_{max}$ - maximal admissible angle, $\nu_1$- width of
      BS for polygonalization, $\nu_2$- width of BS for collinear test[1]
Result: ARC (resp. CIRCLE): $C$ is a digital arc (resp. circle), FALSE if not.
begin
    Use algorithm [13] to polygonalize $P$ into BS of width $\nu_1$: $C = \{C\}_{i=0}^{m}$;
    Represent $C$ in the modified tangent space by $T(C)$; $BS = \emptyset$;
    if there exists $i$ such that $|T_{i2}.y - T_{i1}.y| > \alpha_{max}$ then return FALSE;
    Determine midpoint set $MpC = \{M_i\}_{i=0}^{m-1}$ of $\{T_{i2}T_{(i+1)1}\}_{i=0}^{m-1}$; $i = 1$;
    while $|M_i.y - M_0.y| \leq 2\pi$ do $BS = BS \cup M_i$; i++;
    Use algorithm [13] to verify if $BS$ is a blurred segment of width $\nu_2$;
    if $BS$ is a blurred segment of with $\nu_2$ then
        if $|M_{m-1}.y - M_0.y| == 2 * \pi$ then return CIRCLE;
        else return ARC;
    else return FALSE;
end
```

A higher value of the range of vertical ordinate in the tangent space leads to false positive detection as an example; an helix can be detected as an arc. So, to avoid this problem, the maximal difference of vertical ordinate is fixed to 2π for detecting an arc. The input parameter α_{max} of Algorithm 1 allows to control the obtained error.Parameter ν_1 is used for polygonalization by using the recognition of blurred segments. The input parameter ν_2 is used as the width in the algorithm for recognition of blurred segments [13] to test the collinearity of the midpoint set in the representation of the tangent space. In practice, α_{max} (resp. ν_1) is chosen as $\frac{\pi}{4}$ (resp. 1). In addition, ν_2 can be chosen as a fixed value from 0.15 to 0.25 without problem.

This heuristic algorithm detect well circular arc. In Section 5 (see corollary 1), we consider an adaptive estimation of ν_2 to guarantee the quality of detected arcs. It is evident that we can lightly change Algorithm 1 to check the condition of Corollary 1 and Proposition 2 in linear time also.

The algorithms of polygonalization and of blurred segment recognition are linear. The complexity of the tangent space transform, and the construction

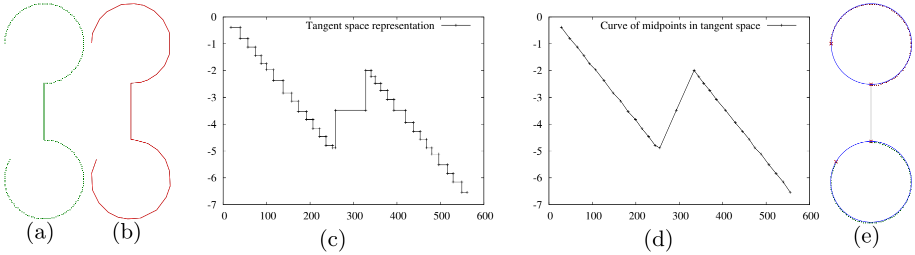
[1] By default, $\alpha_{max} = \frac{\pi}{4}$, $\nu_1 = 1$ for normal curves and $\nu_2 = 0.2$ (see algo. 1).

of midpoint curve $MpC = \{M_i\}_{i=1}^{m-1}$ is in $O(m)$. Because $m << n$, the total complexity of our detection method is then in $O(n)$.

Segmentation of curves into digital arcs. Based on the above idea for detection of an arc, we then develop a linear method for the segmentation of digital arcs by using a width ν blurred segment [13] polygonalization on the curve of midpoints. Its main idea, illustrated in figure 5, is based on the polygonalization of the midpoint curve (Fig. 5.e). Contrariwise to Algorithm 1, we polygonalize the midpoint curve MpC in spite of recognizing if MpC is a BS and then each line segment corresponds to a circular arc.

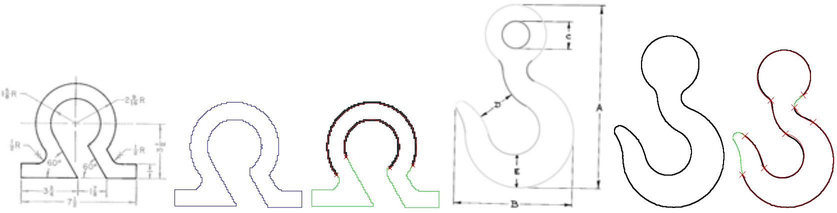
Experimental results and application to real images. We have implemented this linear method. An example of a curve segmented into arcs is presented in Fig. 5. Firstly, the approximating polygon (see Fig. 5.b) is constructed from the input curve in Fig. 5.a. After that, we transform it into the modified tangent space representation (see Fig. 5.c). Then, by polygonalizing the curve of midpoints in this tangent space (see Fig. 5.d), the corresponding arcs can be detected (see Fig. 5.e).

Fig. 6 shows an experimentation on technical drawing images. Figs. 6.a, 6.d are input images. Figs. 6.c, 6.f present the extracted arcs from the borders presented in 6.b, 6.e. Our method gives good results on this type of images which frequently contain arc and circle primitives. Fig. 7 presents our obtained result with a real image.



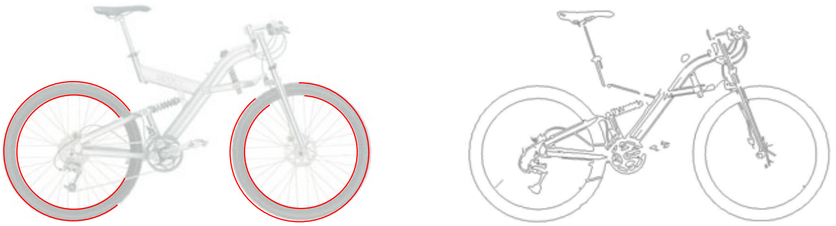
(a) (b) (c) (d) (e)

Fig. 5. Arc segmentation: (a) Input curve, (b) Approximated polygon, (c) Tangent space representation, (d) Curve of midpoints, (e) Results of arc segmentation



(a) Input image (b) Outline (c) Result (d) Input image (e) Outline (f) Result

Fig. 6. Experimentation on technical drawing images



(a) Detected arcs on the input image

(b) Extracted edge using Canny filter

Fig. 7. Experimentation on a real image at width 2

5 Quasi Collinearity Property of Midpoint Curve

Algorithm 1 works well in practice. However, we have a problem to estimate error approximation in general case when the value of ν_2 is fixed. In this section, we present a first study concerning the utilization of this algorithm where ν_2 is chosen adaptively.

Let us suppose that $\alpha_{max} = max\{\alpha_i\}_{i=1}^n$. Let us suppose that R_i is the radius of the approximating circle that passes through 3 points C_{i-1}, C_i, C_{i+1}; $\alpha_{i1} = \angle H_{i-1}OC_i$, $\alpha_{i2} = \angle H_iOC_i$ (see Fig. 3). We suppose that $\alpha_{i1}, \alpha_{i2} \leq \frac{\pi}{8}$ for $i = 1, \ldots, n-1$ to guarantee the condition $\sin x \simeq x$ in Theorem 1. It means that we consider the condition $\sin x \simeq x$ with $x \in [0, \frac{\pi}{8}]$. Therefore, we have $\alpha_i \leq \frac{\pi}{4}$.

Comparison of radius of local circumcircles to the global radius

Proposition 1. *Let $C = \{C_i\}_{i=0}^n$ be a polygon, $\alpha_i = \angle(\overrightarrow{C_{i-1}C_i}, \overrightarrow{C_iC_{i+1}})$. The length of C_iC_{i+1} is l_i, for $i \in \{0, \ldots, n-1\}$. We denote O_i (resp. R_i) respectively the center (resp. the radius) of circumcirle that passes through 3 points C_{i-1}, C_i, C_{i+1}, H_i the projection of O_i on C_iC_{i+1}. Suppose that $R_i - OH_i \leq h$ for $i \in \{1, \ldots, n-1\}$. This results below is obtained. $R_i\alpha_i \geq \frac{l_{i-1}+l_i}{2} \geq R_i\alpha_i - 0.3377h\alpha_i$*

Proof. We denote $\alpha_{i1} = \angle H_{i-1}O_iC_i$, $\alpha_{i2} = \angle H_iO_iC_i$ (see Fig. 3). Firstly, $\cos\alpha_{i1} = 1 - 2\sin^2\frac{\alpha_{i1}}{2} = \frac{OH_i}{R_i} \geq \frac{R_i - h}{R_i} = 1 - \frac{h}{R_i}$. In addition, thanks to $\alpha_{i1} \leq \frac{\pi}{8}$ and $\frac{\sin(x)}{x}$ is decreasing in $[0, \frac{\pi}{16}]$, we have $\sin x \geq x\frac{\sin\frac{\pi}{16}}{\frac{\pi}{16}}$. Therefore $\frac{h}{R_i} \geq 2\sin^2\frac{\alpha_i}{2} \geq 2 \cdot \left(\frac{\sin\frac{\pi}{16}}{\frac{\pi}{16}}\right)^2 \left(\frac{\alpha_{i1}}{2}\right)^2 > \frac{0.9872}{2}\alpha_{i1}^2$. Similarly, we have $\frac{h}{R_i} > \frac{0.9872}{2}\alpha_{i2}^2$, for $i \in \{1, \ldots, n-1\}$ (1).

In addition, we have this remark $x \geq \sin x \geq x - \frac{x^3}{6}$ with $\frac{\pi}{4} \geq x \geq 0$. So, $\alpha_i \geq \sin\alpha_{i1} + \sin\alpha_{i2} > \alpha_{i1} + \alpha_{i2} - \frac{1}{6}(\alpha_{i1}^3 + \alpha_{i2}^3) = (\alpha_{i1} + \alpha_{i2})(1 - \frac{1}{6}(\alpha_{i1}^2 - \alpha_{i1}\alpha_{i2} + \alpha_{i2}^2)) = \alpha_i(1 - \frac{1}{6}(\alpha_{i1}^2 - \alpha_{i1}\alpha_{i2} + \alpha_{i2}^2)) \geq \alpha_i(1 - \frac{1}{6}(\alpha_{i1}^2 + \alpha_{i2}^2)))$, for $i \in \{1, \ldots, n-1\}$ (2).

Thanks to (1) and (2), we obtain $\sin\alpha_{i1} + \sin\alpha_{i2} > \alpha_1(1 - \frac{1}{3 \cdot 0.9872}\frac{h}{R})$, for $i \in \{1, \ldots, n-1\}$ (3).

Moreover, we have $\alpha_i = \alpha_{i1} + \alpha_{i2}$ and $\alpha_{i1}, \alpha_{i2} \leq \frac{\pi}{8}$. In addition, we have $\sin\alpha_{i1} = \frac{l_{i-1}}{2R}$, $\sin\alpha_{i2} = \frac{l_i}{2R}$. Therefore, we have $\frac{l_{i-1}+l_i}{2R} = \sin\alpha_{i1} + \sin\alpha_{i2}$, for $1 \leq i < n$ (4).

Thanks to (3) and (4), we obtain $R_i\alpha_i \geq \frac{l_{i-1}+l_i}{2} \geq R_i\alpha_i(1 - \frac{1}{3 \cdot 0.9872}\frac{h}{R_i}) = \alpha_i(R_i - \frac{h}{3 \cdot 0.9872}) \Leftrightarrow R_i\alpha_i \geq \frac{l_{i-1}+l_i}{2} \geq R_i\alpha_i - 0.3377h\alpha_i$.

Now we consider a set of midpoints MpC is a blurred segment whose horizontal width is ϵ. Let us suppose that the slope of this blurred segment is $\frac{1}{R}, R \in \mathbb{R}$.

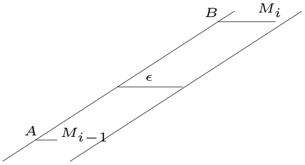


Fig. 8.

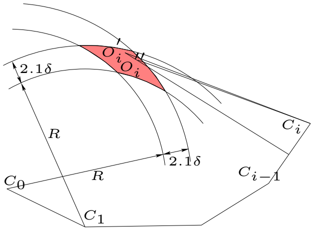


Fig. 9. $\{O_i^{'}, O_i^{''}\}_{i=1}^{n-1}$ is in a compact zone

Let us consider Fig. 8, A and B respectively are horizontal projection of M_{i-1} and M_i on the left leaning line. We have: $M_i.x - M_{i-1}.x = (B.x - A.x) + M_iB - AM_{i-1} = R\alpha_i + M_iB - AM_{i-1}$. Because of M_i and M_{i-1} are limited by 2 leaning lines, we have $M_iB \leq \epsilon$ and $AM_{i-1} \leq \epsilon$, so $-\epsilon \leq M_iB - AM_{i-1} \leq \epsilon$. Therefore, we have $R\alpha_i + \epsilon \geq \frac{l_{i-1}+l_i}{2} \geq R\alpha_i - \epsilon$. This double inequations can be rewrited as $\frac{l_{i-1}+l_i}{2} = R\alpha_i + \epsilon_i$, where $\epsilon_i \in [-\epsilon, \epsilon]$.

Thank to proposition 1, we have: $R_i\alpha_i > R\alpha_i + \epsilon_i > R_i\alpha_i - 0.3377h\alpha_i \Leftrightarrow R_i > R + \frac{\epsilon_i}{\alpha_i} > R_i - 0.3377h$. So, $\frac{\epsilon}{\alpha_i} < R_i - R < \frac{\epsilon}{\alpha_i} + 0.3377h$. So, we have corollary 1.

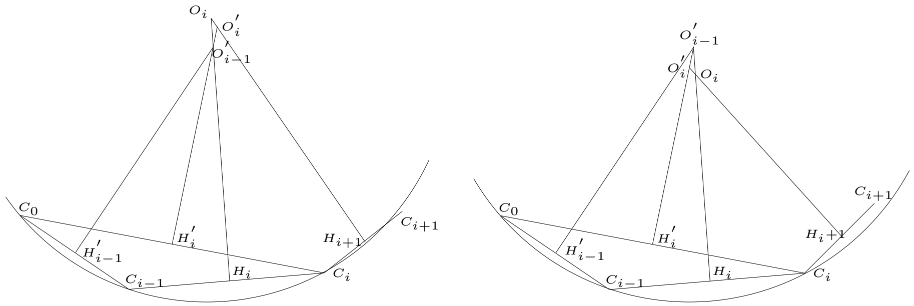
Corollary 1. *Let $C = \{C_i\}_{i=0}^n$ be a polygon, $\alpha_i = \angle(\overrightarrow{C_{i-1}C_i}, \overrightarrow{C_iC_{i+1}})$. The length of C_iC_{i+1} is l_i, for $i \in \{0, \ldots, n-1\}$. The set of midpoints $\{M_i\}_{i=0}^{n-1}$ is a blurred segment whose horizontal width is ϵ. We denote O_i (resp. R_i) respectively the center (resp. the radius) of circumcirle that passes through 3 points C_{i-1}, C_i, C_{i+1}, H_i the projection of O_i on C_iC_{i+1}. Suppose that $R_i - OH_i \leq h$ for $i \in \{1, \ldots, n-1\}$. This results below is obtained.* $0 < R_i - R < \frac{\epsilon}{\alpha_i} + 0.3377h \leq \frac{\epsilon}{\min\{\alpha_i\}_{i=1}^n} + 0.3377h.$

Localization of centers of local circumcircles. In this section, we consider the convergence of local circumcircle centers if this condition below is satisfied: $|R_i - R| \leq \delta$, $R \in \mathbb{R}$, $1 \leq i, j \leq n-1$. Let us consider Fig. 10. To prove the convergence of centers of local circumcircle, we show this property for an approximation of a half of circle.

Proposition 2. *Let us consider a sequence of points* $\{C\}_{i=0}^{n}$*. There exist* R *and* δ *such that* $R, \delta \in \mathcal{R}, 0 \leq R_i - R \leq \delta, i = 1, \ldots, n-1$*. Suppose that* $\angle C_k C_j C_{j+1} > \frac{\pi}{2}$ *for* $k \in \{0, 1\}, k < j < n$*. Therefore, we have this property* $0 \leq R_i^{'} - R \leq \delta$*,* $0 \leq R_i^{''} - R \leq \delta$*, for* $1 \leq i \leq n-1$*.*

Proof. We denote $O_i^{'}$ (resp. $O_i^{''}$) and $R_i^{'}$ (resp. $R_i^{''}$) the centers and radius of circumcircles that passes through 3 points C_0 (resp. C_1), C_i, C_{i+1}. Firstly, we have a trivial remark: the perpendicular bisector of C_iC_k is between that of C_iC_j and C_jC_k. Now, we prove this proposition by induction.

Because of $R_1^{'} = R_1$, the proposition is true with $i = 1$. Suppose that $|R_{i-1}^{'} - R| \leq \delta$. Let us consider Fig. 10. We denote $H_i^{'}$, H_i are respectively the midpoint of C_0C_i, $C_{i-1}C_i$. Thank to the above remark, we have $O_{i-1}^{'}H_i^{'}$ is between $O_{i-1}^{'}H_{i-1}^{'}$ and $O_{i-1}^{'}H_i$. We consider now the position of C_{i+1} with the circle of center $O_{i-1}^{'}$, of radius $R_{i-1}^{'}$. The proposition if trivial if C_{i+1} is on this circle because of $R_i^{'} = R_{i-1}^{'}$. If C_{i+1} is outside of this circle (see Fig. 10.a), we then deduce $O_{i-1}^{'} \in [O_i^{'}H_i^{'}]$, $O_i^{'} \in [O_iH_{i+1}]$. Therefore, we have $O_{i+1}C_i > O_i^{'}C_i$, $O_i^{'}C_i > O_{i-1}^{'}C_i$. It means that $R_{i+1} > R_i^{'} > R_{i-1}^{'}$. Thank to $0 \leq R_{i+1} - R \leq \delta$ and $0 \leq R_{i-1}^{'} - R \leq \delta$, we have $0 \leq R_i^{'} - R \leq \delta$. In other case (see Fig. 10.b), by applying the same arguments, we obtain $R_{i+1} < R_i^{'} < R_{i-1}^{'}$. Therefore, we have $0 \leq R_i^{'} - R \leq \delta$, for $1 \leq i \leq n-1$. By replacing C_0 by C_1 and using the same argument, we have $0 \leq R_i^{''} - R \leq \delta$, for $1 \leq i \leq n-1$.



(a) Outside

(b) Inside

Fig. 10. Position between C_{i+1} and circumcircle of $C_0C_{i-1}C_i$

We have a simple remark: a triangle ABC that satisfies $\angle ABC \geq \frac{7\pi}{8}$ have : $AC > BC + \cos\frac{\pi}{8} \cdot AB$. It is trivial because of $AC^2 = BC^2 + AB^2 - 2BC \cdot AB \cos\angle ABC > BC^2 + \cos^2\frac{\pi}{8}AB^2 + 2BC \cdot AB\cos\frac{\pi}{8}$. Therefore, we have: $O_i^{'}O_i^{''} \leq \frac{|R_i^{'} - R_i^{''}|}{\cos\frac{\pi}{8}} \leq \frac{\delta}{\cos\frac{\pi}{8}} \leq 1.1\delta$, for $i = 1, \ldots, n-1$. Thanks to this result and proposition 2, it is trivial now to show that set of center $O_i^{'}$, $O_i^{''}$ is in a compact zone (see Fig. 9).

6 Conclusions

We have presented a linear method for the detection of digital circles or digital arcs. A linear method for the segmentation of a curve into digital arcs is also proposed. This method is based on a discrete geometry approach. It is simple, easy and robust to implement. A more complete demonstration of the algorithm is under process.

References

1. Latecki, L., Lakamper, R.: Shape similarity measure based on correspondence of visual parts. PAMI 22, 1185–1190 (2000)
2. Nakamura, A., Aizawa, K.: Digital circles. Computer Vision, Graphics, and Image Processing 26, 242–255 (1984)
3. Kim, C.E.: Digital disks. PAMI 6, 372–374 (1984)
4. Kim, C.E., Anderson, T.A.: Digital disks and a digital compactness measure. In: STOC, pp. 117–124. ACM, New York (1984)
5. Kovalevsky, V.: New definition and fast recognition of digital straight segments and arcs. In: ICPR, vol. 2, pp. 31–34 (1990)
6. Fisk, S.: Separating point sets by circles, and the recognition of digital disks. PAMI 8, 554–556 (1986)
7. Coeurjolly, D., Gérard, Y., Reveillès, J.P., Tougne, L.: An elementary algorithm for digital arc segmentation. Discrete Applied Mathematics 139, 31–50 (2004)
8. Sauer, P.: On the recognition of digital circles in linear time. Comput. Geom. Theory Appl. 2, 287–302 (1993)
9. Damaschke, P.: The linear time recognition of digital arcs. Pattern Recognition Letters 16, 543–548 (1995)
10. Megiddo, N.: Linear programming in linear time when the dimension is fixed. Journal of the ACM 31, 114–127 (1984)
11. Worring, M., Smeulders, A.: Digitized circular arcs: characterization and parameter estimation. PAMI 17, 587–598 (1995)
12. Roussillon, T., Tougne, L., Sivignon, I.: On three constrained versions of the digital circular arc recognition problem. In: Brlek, S., Reutenauer, C., Provençal, X. (eds.) DGCI 2009. LNCS, vol. 5810, pp. 34–45. Springer, Heidelberg (2009)
13. Debled-Rennesson, I., Feschet, F., Rouyer-Degli, J.: Optimal blurred segments decomposition of noisy shapes in linear time. Computers & Graphics 30 (2006)
14. Andres, E.: Discrete circles, rings and spheres. Computers & Graphics 18, 695–706 (1994)

Multi-cue-Based Crowd Segmentation in Stereo Vision

Ya-Li Hou and Grantham K.H. Pang

Industrial Automation Research Laboratory, Department of Electrical and Electronic
Engineering, The University of Hong Kong, Hong Kong
ylhou@eee.hku.hk, gpang@eee.hku.hk

Abstract. People counting and human detection have always been important objectives in visual surveillance. With the decrease in the cost of stereo cameras, they can potentially be used to develop new algorithms and achieve better accuracy. This paper introduces a multi-cue-based method for individual person segmentation in stereo vision. Shape cues inside the crowd are explored with a block-based Implicit Shape Model. Depth cues are obtained from the disparity values of some foreground blobs, which are calculated concurrently during crowd segmentation. Crowd segmentation is therefore achieved with evidences from both shape and depth cues. The methods were evaluated on two video sequences. The results show that the segmentation performance has been improved when depth cues are considered.

Keywords: Stereo vision, Crowd segmentation, Block-based Implicit Shape Model, Disparity.

1 Introduction

Various techniques for segmenting individual pedestrians have been investigated based on monocular camera. As the cost of stereo cameras decreases, interests in people detection based on stereo vision have been increasing. Some researchers [1, 2] perform disparity calculation algorithm first and then followed by human detection in the 2D image space. In [1], the foreground is segmented into multiple blobs based on different disparity values. Human detection methods based on appearance are performed to help grouping and dividing the blobs. In [2], 2D image based human detection is performed in each image independently. The results are verified based on the epipolar geometry.

When the calibration parameters are available, many researchers reconstruct the scene in the 3D space. To handle the occlusion situation, a virtual camera is assumed to get the plan-view map of the 3D points. Different plan-view statistics, like occupancy map, height map have been attempted for human detection and tracking [3, 4]. Kelly et al. [5] proposed a region clustering algorithm to achieve better robustness to occlusion situations. In their later work [6], a more detailed biometric human model is used to avoid over- or under- segmentation during the clustering process. These methods rely on the recovery of the 3D points. However, when people are far away from the camera, the depth information is difficult to obtain.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 93–101, 2011.

Most of the reviewed methods use a disparity algorithm as the pre-processing step of the detection method. Commonly used disparity algorithms include three categories. Correlation-based algorithms calculate the disparity for each local region independently. It does not consider the context among the local regions. Hence, the results are usually quite noisy, which is difficult to be used for crowd segmentation. Dynamic programming based algorithms find the globally optimal correspondence point pairs line by line. Kelly et al. [6] tried to explore the scene features in surveillance videos to achieve more reliable disparity values. However, the assumption of the independence between lines usually produces artifacts along the lines. The third category is area based algorithms, like graph cut. This category assumes that regions with similar color or texture should have similar disparity values. These methods are computationally expensive and errors happen when two close objects show similar color.

All the disparity algorithms have an assumption based on low-level image features, which may be incorrect in many situations. In addition, the noisy disparity values even inside one person make the individual segmentation difficult. In this paper, disparity value calculation and crowd segmentation are performed concurrently. This method has several advantages: First, the method assumes that body parts of one person have the same depth. The assumption is made on a semantic level and it is reasonable when the camera is far away from the camera compared with human depth intra-variation. Second, the developed method would only calculate one disparity value for each person candidate, which would make the crowd segmentation process easy. In this paper, the requisites of the stereo camera setup are as follows. First, the two cameras in the stereo setup have a short baseline distance. In this way, scenes in left image and right image would have a large overlap. Second, objects are far away from the camera compared with human depth intra-variation. In this way, the body parts of a person can be assumed to have the same disparity value.

2 System Overview

Fig. 1 shows a block diagram of the proposed system. Shape and depth cues are considered concurrently. Shape cues are collected with a Block-based Implicit Shape Model and depth cues are calculated in a semantic level. The key steps will be introduced in details in the following sections.

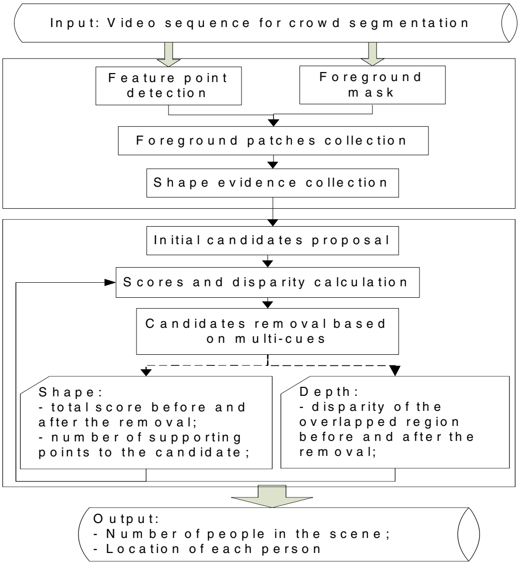

Input: Video sequence for crowd segmentation
Feature point detection
Foreground mask
Foreground patches collection
Shape evidence collection
Initial candidates proposal
Scores and disparity calculation
Candidates removal based on multi-cues
Shape:
- total score before and after the removal;
- number of supporting points to the candidate;
Depth:
- disparity of the overlapped region before and after the removal;
Output:
- Number of people in the scene;
- Location of each person

Fig. 1. Block diagram of the method

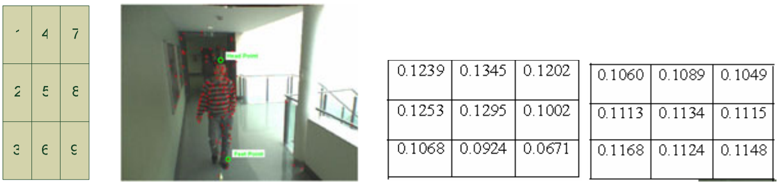
2.1 Initialization

The initialization step is similar as [7]. Since our later steps are built upon it, we will briefly review the main ideas.

Before the segmentation, a training stage is necessary to establish a human shape model. In our system, the Block-based Implicit Shape Model (B-ISM) is used. A KLT feature point detector [8] is applied on the training persons. Then, a number of training patches are collected around the feature points and grouped into several clusters based on their shape descriptor - Histogram of Oriented Gradients (HOG)[9]. Finally, all the patches vote for the spatial occurrence of each cluster based on their location in the 3x3 blocks (Fig. 2a).

In the testing stage, test patches are extracted around the KLT feature points in the foreground area. The spatial occurrence probabilities of each test patch in the 3x3 blocks are obtained based on the established B-ISM. Each cluster votes for the test patch based on the similarity between the test patch and the cluster centers. As shown in Fig. 2b, the head point has got a higher probability in block-4 while the feet point has a higher probability in the bottom row.

For points with a higher probability in block-1, 4 or 7, an initial rectangle candidate would be formed with the point as the center of the top border. In the evaluations, a very conservative threshold, 0.112 ($\approx 1/9$) is used to make sure that the correct rectangles are formed. Usually, a great number of initial rectangles may be formed for a crowd.



0.1239	0.1345	0.1202
0.1253	0.1295	0.1002
0.1068	0.0924	0.0671

0.1060	0.1089	0.1049
0.1113	0.1134	0.1115
0.1168	0.1124	0.1148

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

Fig. 2. Each KLT point has obtained a 3x3 spatial occurrence table. Usually, the head point has a higher probability in top row while the feet point has a higher probability in the bottom rows. (a) 3x3 blocks; (b) an example of head and feet points; (c) the spatial occurrence table of the head point; (d) the spatial occurrence table of the feet point.

2.2 Scores and Disparity Calculation

Given a set of rectangle candidates, the test patches are assigned to the rectangles. It is assumed that rectangles with lower y-coordinates are occluded by those with higher y-coordinates. In Fig. 3a, the blue rectangle is occluded by the green one. Based on its location in the associated rectangles, each patch gets a score with (1). i , k , l are the index for blocks, rectangles and test patches. p_{li} is the probability that the patch occur in the associated block, which has been obtained in section 2.1. $\rho(i,k,l)=1$ only when the patch-l falls in the block-i of rectangle-k, otherwise, $\rho(i,k,l)=0$.

$$s_l = \sum_{i=1:9} (p_{li}\rho(i,k,l)) \tag{1}$$

At the same time, the foreground areas are assigned to the rectangles. For each rectangle, the associated foreground regions are shifted to find the best match in the right image. Since the two images from the stereo camera have been rectified, epipolar lines are along the scan-lines. As shown in equation (2), the disparity value is defined as the shift with the maximum number of matched foreground image pixels. In our evaluations, an image pixel is matched when the intensity difference between the two images is below a threshold, *th*. $F^L_{y,x}$ =1 means that the pixel (y, x) is a foreground pixel in the left image and $F^R_{y,x-d}$ =1 means that the shifted pixel in the right image is in the foreground region. Fig. 3c shows the foreground image patches falling in the green rectangle. Fig. 3d shows the foreground image patches within the entire shift range, $d_{\min} \sim d_{\max}$, in the right image. In this way, each rectangle will get one disparity value.

$$d = \max_{d=d_{\min}:d_{\max}} \Big(\sum_{\substack{F^L_{y,x}, F^R_{y,x-d}=1, \\ y,x \in r_k}} ((I^L_{y,x} - I^R_{y,x-d}) < th)\Big) \tag{2}$$

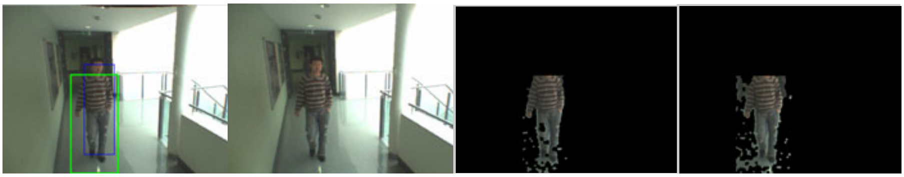

(a) (b) (c) (d)

Fig. 3. Disparity calculation. Disparity value is the shift which has the maximum number of matched foreground pixels. (a) proposed candidates in the left image; (b) the right image; (c) foreground area within the green rectangle in the left image; (d) foreground area within the shift range (0~30) in the right image.

2.3 Removal and Merge

In this step, the initial set of rectangle candidates are examined based on the shape cues and depth cues. Those with insufficient evidences from shape and depth cues will be removed or merged.

Shape cues. Based on the patch scores obtained from last step, candidates without sufficient evidences from shape cues will be removed. In our evaluations, two criteria are considered. First, the total score for the entire crowd configuration is defined as the summation of the all the patch scores. If the removal of the candidate will result in the increase in the total score, it will be removed. That is because a higher total score means that most patches have been allocated to a better block in the crowd configuration. Second, the points which get lower scores after the candidate removal are called the supporting points for the candidate. The existence of the candidate is supported by those supporting points. Hence, if there are few support points for the candidate, it will be removed.

Depth cues. Usually, persons with lower y-coordinates in the image are farther away from the cameras and smaller disparity values are expected. When two candidates with different y-coordinates show similar disparity values, they should be merged into one candidate. An example of the case has been illustrated in Fig. 4. As shown in Fig. 4a, two initial rectangles are proposed for the person. However, similar disparity values have been obtained for the visible parts of the two candidates. The disparity values are indicated with green numbers in Fig. 4b. In Fig. 4c, a new rectangle is proposed based on the merged foreground areas. With the new rectangle, the person can be more accurately located. Finally, the disparity value of the new rectangle is obtained based on the foreground region in it, see Fig. 4d.

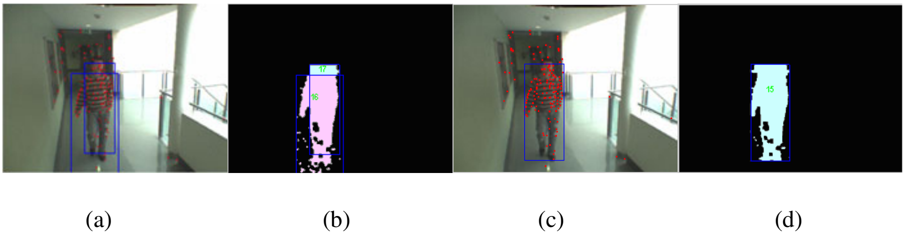


Fig. 4. Multiple candidates are merged due to similar disparity values. (a) multiple initial candidates for one person; (b) disparity calculation for the two candidates respectively; (c) the new rectangle; (d) disparity value of the new rectangle.

After each removal and mergence step, patch scores and disparity calculation are recalculated. The examination process is iteratively performed until the candidates are not changed any more. To get more reliable disparity values, only shape cues are used in the first loop to remove those very close candidates.

3 Evaluations and Results

Datasets. The system was evaluated on two video sets from [10]. The baseline between the two cameras is around 100 mm. The 'corridor' scene was taken by a camera positioned above 2 meters from the ground. With the first 100 frames as the training set, the test was performed on the remaining frames. Starting from the 110th frame, one image was used in every consecutive five frames, that is, 113 testing images in total. In the 'vicon' scene, with the first 80 frames as the training set, the test was performed on the remaining frames. One image was used for testing in every consecutive five frames, that is, 60 testing images in total.

Evaluations. The 2D ground truth is manually obtained. A rectangle is annotated as long as the head is visible in the scene. In the 'corridor' scene, people with y-coordinates smaller than a threshold value would not be counted since they show very few feature points. People on stairs are also excluded because they are not on the same ground plane as others, which makes it difficult to estimate the initial human size.

A detection which has a large overlap (> 50%) with the ground truth is defined as a correct detection. Each ground truth can have only one correct detection. The detected rectangle without a corresponding person is a false detection. Detection rate is defined as *Detection rate=#(correct detection)/#(ground truth)* and false alarm rate is calculated as *False alarm rate=#(false detection)/#(ground truth).*

Results. A fixed background image has been provided in the dataset. To ease the effects of uneven illumination, the foreground region is extracted based on the Hue component in the HSV space in our evaluations. To get a solid foreground mask, a series of morphology operation are performed. An open operation with a circular structuring element is performed to remove scattered noises. After that, a closing operation is performed to form a solid foreground area. To include almost all the feature points from the persons, a dilation operation is also performed. Fig. 5 has shown an example image and the final foreground mask.

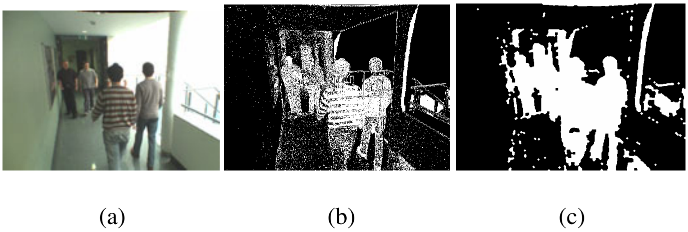
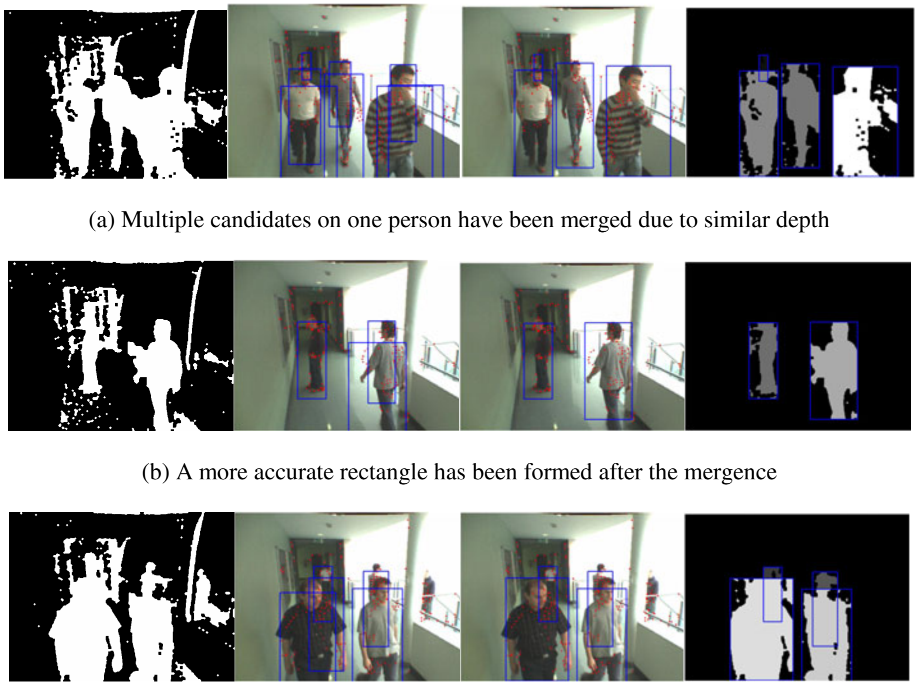


Fig. 5. (a) the original image; (b) foreground image based on the Hue component; (c) foreground mask after a series of morphology operations

Some sample frames on the ‘corridor’ sequence have been shown in Fig. 6. The method based on only shape cues can give some satisfactory results even with a very rough foreground mask. However, when the crowd is dense, or many feature points come from the details, the reliability of the shape cues may become less reliable. In addition, the minimum number of supporting points for a fully-visible person will be difficult to determine when the clothes have many details.

Based on the depth values, candidates with different y-coordinates but similar disparity values are merged. As shown in the third column in Fig. 6, most false detections have been merged and better detections are obtained for the persons. On the other hand, candidates with different disparity values would not be removed even if only small portion is visible, as shown in Fig. 6c. The fourth column shows the disparity values of the final detected persons. The disparity values of the persons in the scene are displayed in grayscale images. Brighter region represents a larger disparity value while darker region represents a lower disparity value. People close to the cameras have got a large disparity value and those further away have a smaller disparity value.

Over all the 113 testing images, the shape-based method has the detection rate of 75.6% and the false alarm rate is 23.9%. After the combination with the depth information, the detection rate is 73.9% while the false alarm rate is 12.6%. The false alarm rate has been significantly reduced while the detection rate remains similar. The miss detections are mainly due to the persons far away from the camera and very low contrast in the dark region. When a person is far away from the cameras, few feature points can be detected and the resolution is low. False detections are mainly from two aspects. First, the method assumes that the entire foreground region is from human beings. A false detection may occur when a large background region is extracted. Second, when the multiple candidates for the same person have similar y-coordinates, the false detection cannot be removed.



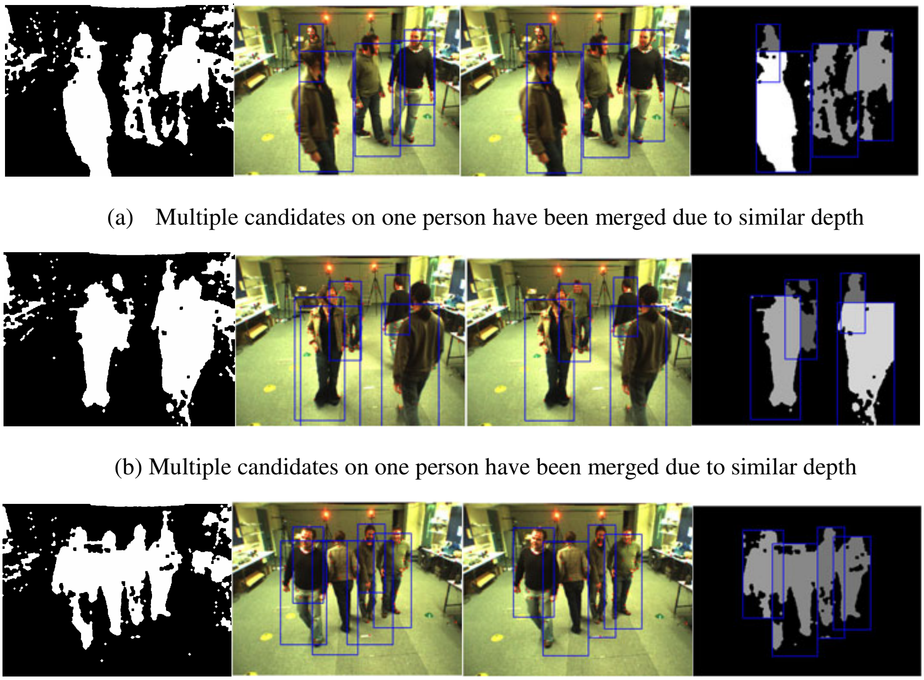
(a) Multiple candidates on one person have been merged due to similar depth

(b) A more accurate rectangle has been formed after the mergence

(c) Seriously occluded persons are not incorrectly merged due to their different depth

Fig. 6. Sample results in the ‘corridor’ sequence. First column: foreground mask; second column: results based on shape cues; third column: results based on shape & depth cues; fourth column: the final disparity map for the persons in the scene.

More results on the ‘vicon-4’ scene have been obtained. With both shape and depth cues, the detection rate is 88% while the false alarm rate is 11.4%. Some sample frames are shown in Fig. 7.



(a) Multiple candidates on one person have been merged due to similar depth

(b) Multiple candidates on one person have been merged due to similar depth

(c) From the final disparity map, different depth of the persons can be observed

Fig. 7. Sample results in the 'vicon-4' sequence. First column: foreground mask; second column: results based on shape cues; third column: results based on shape & depth cues; fourth column: the final disparity map for the persons in the scene.

4 Conclusions

A multi-cue based crowd segmentation method in stereo vision has been introduced in this paper. The contributions are two-fold. First, different from previous methods, the disparity values are calculated concurrently with human detection. It is based on a semantic-level assumption that body parts of a person have similar disparity values. Second, with only one disparity value for each candidate, the depth cue is easier to be used for individual human segmentation. Evaluations show that the use of the depth cue has efficiently reduced the number of false detections.

In the future, a more accurate human model will be used. The accurate model will benefit both shape-based segmentation and disparity calculation.

References

1. Zhao, L., Thorpe, C.E.: Stereo- and neural network-based pedestrian detection. IEEE Transactions on Intelligent Transportation Systems 01, 148–154 (2000)
2. Benezeth, Y., Jodoin, P.M., Emile, B., Laurent, H., Rosenberger, C.: Human Detection with a Multi-sensors Stereovision System. In: Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D., Meunier, J. (eds.) ICISP 2010. LNCS, vol. 6134, pp. 228–235. Springer, Heidelberg (2010)

3. Harville, M.: Stereo person tracking with adaptive plan-view templates of height and occupancy statistics. Image and Vision Computing 22, 127–142 (2004)
4. Muoz-Salinas, R., Aguirre, E., Garca-Silvente, M.: People detection and tracking using stereo vision and color. Image Vision Comput. 25, 995–1007 (2007)
5. Kelly, P., O'Connor, N.E., Smeaton, A.F.: Pedestrian detection in uncontrolled environments using stereo and biometric information. In: Proceedings of the 4th ACM International Workshop on Video Surveillance and Sensor Networks. ACM, Santa Barbara (2006)
6. Kelly, P., Noel, E.O.C., Alan, F.S.: Robust pedestrian detection and tracking in crowded scenes. Image Vision Comput. 27, 1445–1458 (2009)
7. Hou, Y.-L., Pang, G.K.H.: Human Detection in Crowded Scenes. In: IEEE International Conference on Image Processing (2010)
8. Birchfield, S.: Source code of the klt feature tracker (2006), `http://www.ces.clemson.edu/~stb/klt/`
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893 (2005)
10. Kelly, P., O'Connor, N.E., Smeaton, A.F.: A Framework for Evaluating Stereo-Based Pedestrian Detection Techniques. IEEE Transactions on Circuits and Systems for Video Technology 18, 1163–1167 (2008)

Semantic Segmentation of Microscopic Images Using a Morphological Hierarchy

Cristian Smochina[1], Vasile Manta[1], and Walter Kropatsch[2]

[1] Faculty of Automatic Control and Computer Engineering,
”Gheorghe Asachi” Technical University of Iasi, Romania
smochina.cristian@yahoo.com, vmanta@cs.tuiasi.ro
[2] Pattern Recognition and Image Processing Group, Institute of Computer Graphics and Algorithms, Vienna University of Technology, Austria
krw@prip.tuwien.ac.at

Abstract. The objective of semantic segmentation in microscopic images is to extract the cellular, nuclear or tissue components. This problem is challenging due to the large variations of these components features (size, shape, orientation or texture). In this paper we present an automatic technique to robustly identify the epithelial nuclei (crypt) against interstitial nuclei in microscopic images taken from colon tissues. The relationship between the histological structures (epithelial layer, lumen and stroma) and the ring like shape of the crypt are considered. The crypt inner boundary is detected using a closing morphological hierarchy and its associated binary hierarchy. The outer border is determined by the epithelial nuclei, overlapped by the maximal isoline of the inner boundary. The evaluation of the proposed method is made by computing the percentage of the mis-segmented nuclei against epithelial nuclei per crypt.

Keywords: Crypt segmentation, Morphological hierarchy, Biomedical imaging, Pathology, Microscopy.

1 Introduction

In diagnostic pathology, the pathologists give a diagnostic after a set of biological samples (tissues stained with different markers) are viewed and many specific features of the objects of interest (size, shape, colour or texture) have been analysed. This complex diagnostic process is an important part in clinical medicine but also in biomedical research and can be enhanced by providing the pathologists or the biologists with quantitative data extracted from the images. The image processing techniques are of special interest because they allow large scale statistical evaluation in addition to classical eye screening evaluation and are used in both sections of the pathology: cytology (the study of cells) and histology (anatomical study of the microscopic structure of tissues) [1]. The microscopic image segmentation is considered a hard task due to the following problems also pointed out in [2]:

- Low contrast and weak boundaries on out-of-focus nuclei. Also some nuclei structures can appear as artefacts in the non-uniformly stained slices.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 102–109, 2011.

- Different grey values for the background cased by the non-uniform illumination.
- The physical structure of the cells and the way of sectioning determine a non-uniform distribution of material inside the nucleus (lower intensities within nuclei).
- Considerable variation of object features like shape and/or size and/or orientation and different nuclei distribution within the epithelial layer.

This paper is organized as follows. The last part of this section points out the goal of this study. The inner boundaries of the crypts are detected using the morphological hierarchy and the lumen reconstruction (section 2), while in section 3 the outer borders are detected using the maximal isoline. The results are evaluated in section 4 and discussed and concluded in section 5.

1.1 State of the Art

Many studies from the literature cover fields like microscopy, biomedical engineering and imaging, bioinformatics or pattern recognition and introduce techniques for solving the mentioned problems [2]. Beside the nuclei segmentation attempts [1], [2], also the segmentation of histological structures like gland or crypt is addressed. In [3] a threshold is used to identify the gland seeds which are grown to obtain the nuclei chain. In [4] the pixel labelling to different classes is performed using a clustering approach based on the textural properties.

An object-graphs approach is described in [5] where the relationship between the primitive objects (nucleus and lumen) is considered. The prostate cancer malignancy is automatically graded (Gleason system) in [6] after the prostate glands are detected.

1.2 Aim of the Study

The basic functional unit of the small intestine is the crypt (crypt of Lieberkühn) [7] and it comprises two main structures of interest: the lumen and the epithelial layer (Fig. 1a). The epithelial layer contains epithelial nuclei and surrounds the lumen which is an empty area. The interstitial cells on the other side form heterogeneous regions (stroma) placed between crypts. The stroma areas (Fig. 1a) contain isolated cells with non-regular shape and without particular patterns of arrangement.

This work provides specific techniques to segment the crypts from fluorescence images of colorectal cancer tissue sections. We used 8 bit greyscale images (Fig. 1a) containing nuclei labelled with DAPI, a fluorescent stain that binds strongly to DNA [8] and acquired using a TissueFAXS slide scanner (TissueGnostics GmbH, Austria).

The main motivation for segmenting crypts is to provide the pathologists with quantitative data regarding the areas covered by epithelial nuclei. One alternative approach is to segment each nucleus and to analyze the structures that they form. Since this approach can encounter additional problems, our objective is to directly find the boundaries of the crypts, i.e. to delineate the area covered by the epithelial nuclei without dealing with the individual nuclei (Fig. 2). In [5] also the high level information are preferred against the *local* one but their approach uses different images type (hematoxylin-and-eosin stained images) which provide more biological details (also cell cytoplasm is available) than our DAPI stained nuclei images.

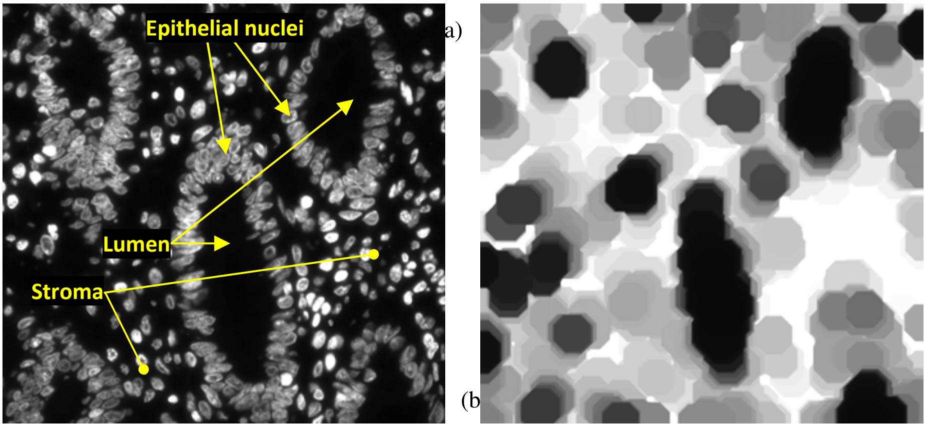


Fig. 1. a) Fluorescence image with crypts from a colon tissue section. b) The image from the top level of the morphological hierarchy. The black regions indicate the lumen.

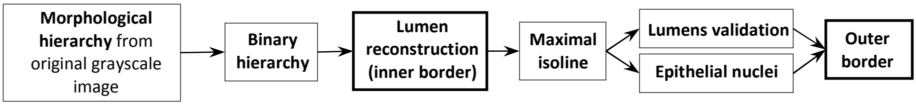

Morphological hierarchy from original grayscale image
Binary hierarchy
Lumen reconstruction (inner border)
Maximal isoline
Lumens validation
Epithelial nuclei
Outer border

Fig. 2. Overview scheme of the proposed technique

2 Lumens Segmentation

Without considering the relations between the crypt components (epithelial layer and lumen), the stroma and the background, the low level cues will not be able to separate the regions having a particular meaning [9]. A way must be found to keep only the important information and to remove the unnecessary details. In order to detect these regions, the role of *local* information (pixel grey values or gradient) is very important but not sufficient; also *global* information like the region's size and relation with the other region types must be included [9].

A rough assumption about the nuclei distribution over different region types can be made. The lumen does not contain nuclei and appears like a big round black area surrounded by a 'ring' with variable thickness. This ring contains a high density of touched/chained epithelial cells. The exceptional cases appear when the lumen gets to be in touch with the stroma area due to missing cells that 'break' this ring.

By applying the morphological closing operation [10] on the grey image, the nuclei closer than the size of the structure element (SE) will be connected. The epithelial nuclei and those from the stroma area can be connected by relating the SE's size to the size of the lumen region (the connection should not pass over the lumen). For that we build the hierarchical image decomposition similar to the morphological pyramid [11]: the upper levels are obtaining by applying a morphological operator on the base image. The difference consists in lack of sub-sampling step [12].

2.1 Building the Morphological Hierarchy

Let I denote the input grey scale image and $\bullet$ denote the morphological closing operation. The SE ψ_k is a two-dimensional disk of diameter $2k+1$. The hierarchical morphological representation $\Pi_\bullet$ consists of L levels. The first one is the original grey scale image $\Pi_\bullet^1 = I$ and each level $\ell > 0$ is given by $\Pi_\bullet^\ell = \Pi_\bullet^1 \bullet \psi_{2\ell}, \ell = \overline{1,L}$.

The closing operation smoothes the objects' boundary and removes the dark holes smaller than the SE. Since the size of the SE increases according to the level of the hierarchy, in the lower levels only the small gaps will be filled, while the bigger ones will be closed in the upper levels. To prevent the SE from growing too large, the maximum number of levels is established by limiting the SE's size so that it covers maximally 2-3 nuclei (in our experiments the SE ψ_{50} gives 25 levels). The lumens should 'survive' till the top level (Fig. 1b) and should be easier highlighted; also the gaps from the crypts (do not exceed the size of 1-2 nuclei) should be filled.

2.2 Lumen Reconstruction

The proper reconstruction of each lumen based on the found regions from the top level must be done by analyzing the lower levels of the hierarchy where more details are present. A binary hierarchy Π_{bw} is build in which each level represents the result of the thresholding applied on the corresponding level from the $\Pi_\bullet$. The hierarchy Π_{bw} consists also of L levels and each level l is given by

$$\Pi_{bw}^l = \Pi_\bullet^l < c \cdot thr_{\text{Otsu}}(\Pi_\bullet^l), l = \overline{1,L} \qquad (1)$$

where $thr_{\text{Otsu}}(\cdot)$ computes the threshold for an image using the Otsu's method [13] and $0 < c \leq 1$ (0.5 in our experiments).

Our goal is to find for each partition (ancestor) from the top of Π_{bw} the corresponding partition (descendent) from a bottom level which properly identifies the lumen. Each level l of the Π_{bw} contains np_l unconnected regions $\Pi_{bw}^l = \{P_1^l, P_2^l, P_3^l, ..., P_{np_l}^l\}$. A vertical relation between partitions of successive levels can be established: each partition of a level is included in a partition from the below level.

$$\forall P_p^l (2 \leq l \leq L, 1 \leq p \leq np_l), \exists p' (1 \leq p' \leq np_{l-1}), \text{such that } P_p^l \subset P_{p'}^{l-1} \qquad (2)$$

This inclusion is valid due to the reduction of the black regions caused by the increasing of SE's size used in the closing operation. According to Eq. 2, for each partition of any level of Π_{bw}, the corresponding partition from the base level can be found so that the inclusion rule is validated.

The base level Π_{bw}^1 will not give for sure the proper regions because the global threshold is applied on the original image where the main structures are not properly highlighted. For each partition of the top level $P_r^L, 1 \leq r \leq np_L$, the corresponding descendent from the base level P_r^1 is obtained. By bottom-up analyzing level by level, a certain level l_r is chosen such that the ancestor region ($P_r^{l_r}$) of P_r^1 from level l_r (having as ancestor the P_r^L) checks the following rules:

- $\text{solid}(P_r^{l_r}) > min\_solidity$. The function solid(·) computes the proportion of the pixels from the convex hull that are also in the region; $min\_solidy = 0.8$ in our experiments.
- $\text{distEuclid}(\text{centroid}(P_r^{l_r}), \text{centroid}(P_r^{l_r - 1})) < max\_dist$. The function centroid(·) returns the centroid of a region and the distEuclid(·) computes the Euclidean distance between these two centroids. This rule ensures stability by checking the distance between the regions centroids from two successive levels.

The border of the found lumens actually describes the inner border of the crypt (Fig. 3a). The false positive (FP) results are eliminated by a validation rule in 3.1.

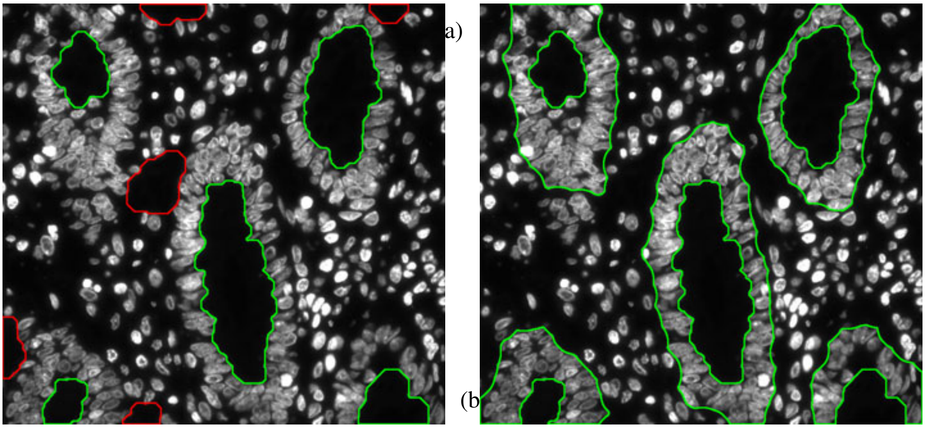


Fig. 3. a) The true positive (green curves) and the FP (red curves) lumen boundaries. b) The green curves indicate the inner and the outer boundaries of the crypts.

3 Crypt’s Outer Border

The epithelial layer is differentiated from the stroma areas by considering the nuclei distribution: the crypt’s nuclei are packed tightly together while those from the stroma areas are wide spread with considerable distances between them. The isolines of the inner boundary are used to eliminate the FP lumens and to detect the outer border which delineates the epithelial nuclei. Each isoline contains pixels situated on the same distance from the inner boundary.

The maximum distance $d_{\max}$ (60 in our experiments) is related to the average width of the epithelial layer. Fig. 4b displays the smoothed signal containing the intensities sum for each isoline. The maximum value (green square) indicates the isoline which gives the maximum sum of pixels intensities. This maximal isoline (depicted with blue in Fig. 4a) is used for two purposes: to validate the detected lumens and to mark the epithelial nuclei (used to get the outer boundary).

3.1 Lumen Validation

The area covered by nuclei can be identified by subtracting a highly blurred version from the original image: $N_{bw} = (I - G * I) > 0$, where G is a big Gaussian filter (201 by 201 in our experiments) and $*$ denotes the convolution operation. Considering the high nuclei concentration around the lumen, there should be only few situations in which the maximal isoline crosses over the background in N_{bw} (Fig. 4a) i.e. situations of big distances between epithelial nuclei or in case of crypt breaks.

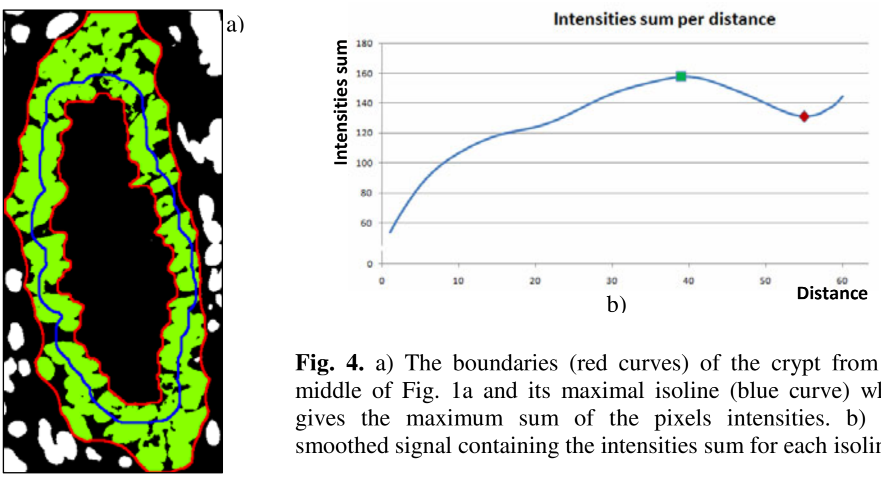


Fig. 4. a) The boundaries (red curves) of the crypt from the middle of Fig. 1a and its maximal isoline (blue curve) which gives the maximum sum of the pixels intensities. b) The smoothed signal containing the intensities sum for each isoline.

Based on this, the following rule is proposed to validate the lumen results from 2.2: if the portions of the maximal isoline overlapping the background in N_{bw} are not considerable high compared to those overlapping the nuclei than the found boundary does not delimit a lumen area.

$$r = \frac{card(MI \cap \sim N_{bw})}{card(MI \cap N_{bw})} \tag{3}$$

The binary image MI contains the maximal isoline, $card(\cdot)$ is the cardinality function and ~ gives the complement of a binary image. The FP boundaries ($r > r_{\min}$) are depicted with red in Fig. 3a and the true positive (TP) inner boundaries ($r \le r_{\min}$) with green ($r_{\min} = 0.3$ in our experiments).

3.2 Outer Border Detection

A region from N_{bw} is marked as part of a crypt iff it is overlapped by the band formed by the inner boundary and the maximal isoline. The epithelial nuclei are depicted with green in Fig. 4a. The outer border of the crypt (Fig. 4a the red curve outward) represents the exterior perimeter of the morphological closing applied on the found epithelial nuclei with a SE covering 2-3 nuclei.

4 Results

We tested the proposed segmentation technique on different datasets of images from tissues labelled with DAPI; some results are show in Fig. 3b and Fig.5. The results confirmed that the proposed method could efficiently segment the crypts with a high degree of accuracy.

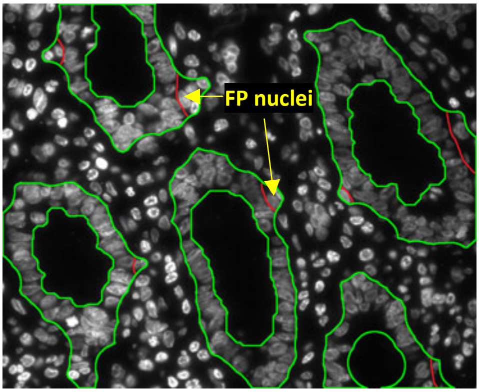


Fig. 5. The green curves indicate the inner and the outer crypt boundaries. The red portions delimit the FP nuclei.

A more rigorous evaluation must be done by comparing the results against the ground truth segmentations. Since a database with reference segmentations for this type of images does not yet exist, a pathology specialist has been asked to validate a set of results. The segmentation quality is established by visual inspecting the number of the mis-segmented nuclei per crypt. A number of 87 crypts have been analyzed resulting in 284 over segmented nuclei. Considering an average of 55 nuclei per crypt, the over-segmented nuclei represent 5.93% from the total crypt's nuclei (an average accuracy of 94.07% per crypt).

5 Conclusions

A new automatic technique for robust crypt segmentation based on hierarchical structures is presented in this paper. A closing morphological hierarchy is used to identify the lumen positions and a binary hierarchy provides enough details for proper lumen reconstruction. The maximal isoline is used to eliminate the false positive lumens and to validate the epithelial nuclei belonging to crypts.

A significant implication of the current work consists of the top-down approach. Firstly the 'obvious' areas (lumens) are detected followed by a more detailed analysis to proper delimit the crypts. The morphological closing operator has been chosen to build the hierarchical representation due to the patterns of nuclei arrangements.

This technique uses a coarser-to-fine approach and can be easily extended on any image with human body cell nuclei from different tissues types (e.g. prostate, breast or lung) but also in any other field in which the objects of interest have the features considered in designing this method. This study will be continued by analysing the topological properties of the graph associated to the tissues components. Considerable effort will be spent to obtain a database with ground-truth segmentations and to find rigorous evaluation criteria of the results.

Acknowledgments. This work was supported in part by the "BRAIN - An Investment in Intelligence" doctoral Scholarship Program within the Technical University of Iasi, Romania and by the Austrian Science Fund under grant no. P20134-N13.

References

1. Ta, V.T., Lezoray, O., El Moataz, A., Schupp, S.: Graph-based tools for microscopic cellular image segmentation. Pattern Recognition 42(6), 1113–1125 (2009)
2. Nattkemper, T.W.: Automatic segmentation of digital micrographs: A survey. Journal Studies in Health Technology and Informatics 107(2), 847–852 (2005)
3. Wu, H.S., Xu, R., Harpaz, N., Burstein, D., Gil, J.: Segmentation of intestinal gland images with iterative region growing. Journal of Microscopy 220(3), 190–204 (2005)
4. Farjam, R., Soltanian-Zadeh, H., Jafari-Khouzani, K., Zoroofi, R.A.: An image analysis approach for automatic malignancy determination of prostate pathological images. Clinical Cytometry 72B(4), 227–240 (2007)
5. Gunduz-Demir, C., Kandemir, M., Tosun, A.B., Sokmensuer, C.: Automatic segmentation of colon glands using object-graphs. Medical Image Analysis 14(1), 1–12 (2010)
6. Naik, S., Doyle, S., Feldman, M., Tomaszewski, J., Madabhushi, A.: Gland segmentation and computerized Gleason grading of prostate histology by integrating low-, high-level and domain specific information. In: 2nd MICCAI Workshop Microscopic Image Analysis with Appl. in Biology, Piscataway, NJ, USA, (2007)
7. Humphries, A., Wright, N.A.: Colonic Crypt Organization and Tumorigenesis: Human Colonic Crypts. Nature Reviews Cancer 8(6), 415–424 (2008)
8. Morikawa, K., Yanagida, J.: Visualization of individual DNA molecules in solution by light microscopy: DAPI staining method. Japanese Biochemical Society 89(2), 693–696 (1981)
9. Kropatsch, W.G., Haxhimusa, Y., Ion, A.: Multiresolution Image Segmentations in Graph Pyramids. In: Kandel, A., Bunke, H., Last, M. (eds.) Applied Graph Theory in Computer Vision and Pattern Recognition, vol. 52, pp. 3–41. Springer, Wien (2007)
10. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 3rd edn. Addison-Wesley, Reading (1992)
11. Haralick, R., Lee, J., Lin, C., Zhuang, X.: Multi-resolution Morphology. In: First IEEE Conference on Computer Vision, London (1987)
12. Haralick, R.M., Zhuang, X., Lin, C., Lee, J.S.J.: The digital morphological sampling theorem. IEEE Trans. Acoust., Speech, Signal Processing 37, 2067–2090 (1989)
13. Otsu, N.: A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics 9(1), 62–66 (1979)

Normalized Joint Mutual Information Measure for Image Segmentation Evaluation with Multiple Ground-Truth Images

Xue Bai, Yibiao Zhao, Yaping Huang, and Siwei Luo

School of Computer and Information Technology, Beijing Jiaotong University
bjtuxbai@gmail.com

Abstract. Supervised or ground-truth-based image segmentation evaluation paradigm plays an important role in objectively evaluating segmentation algorithms. So far, many evaluation methods in terms of comparing clusterings in machine learning field have been developed. Being different from recognition task, image segmentation is considered an ill-defined problem. In a hand-labeled segmentations dataset, for the same image, different human subjects always produce various segmented results, leading to more than one ground-truth segmentations for an image. Thus, it is necessary to extend the traditional pairwise similarity measures that compare a machine generated clustering and a "true" clustering to handle multiple ground-truth clusterings. In this paper, based on the Normalized Mutual Information (NMI) which is a popular information theoretic measure for clustering comparison, we propose to utilize the Normalized Joint Mutual Information (NJMI), an extension of the NMI, to achieve the goal mentioned above. We illustrate the effectiveness of NJMI for objective segmentation evaluation with multiple ground-truth segmentations by testing it on images from Berkeley segmentation dataset.

Keywords: image segmentation evaluation, similarity measure, joint mutual information.

1 Introduction

Image segmentation is an indispensable pre-processing step in many vision systems. Many efforts have been devoted to developing more effective segmentation techniques, as well as quantifying the performance of current algorithms. However, due to the ill-defined nature of the segmentation problem, evaluation for segmentation results is still a challenging task. In order to obtain more objective evaluation scores instead of just using subjective judgments, a database of human segmented natural images [1] was established. Therefore, based on the "true" segmentations, ground-truth-based (GT-based) evaluation paradigm is preferred. In this paradigm, most evaluation methods can be either region-based or boundary-based. Here, we focus on the methods used for evaluating region-based segmentation algorithms.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 110–117, 2011.

Since region segmentation can be seen as a clustering procedure for image pixels according to the feature vector for each pixel including color and spacial information, it is a natural way to do the evaluation task in terms of clusterings comparison, i.e. compare the machine outputs against the ground-truth segmentations through some measure of similarity. So far, a lot of clustering-comparison measures have been proposed in machine learning domain, and they can be categorized into three classes which are pair-counting based (e.g. Rand Index [2]), set-matching based (e.g. $\mathcal{H}$ criterion [3]), and information theoretic based similarity measures (e.g. Normalized Mutual Information [4] and Variation of Information [5]). In [6], various clustering-comparison measures are applied to GT-based segmentation evaluation, based on both range images and intensity images, and the experimental results demonstrate their usefulness and applicability in quantifying the performance of segmentation algorithms.

In practice, there is always a set of manual segmentations for each image in a hand-labeled dataset, as different human subjects would produce different segmented results at various granularity levels. So, the similarity measures should be extended to deal with multiple ground-truth images. Unnikrishnan et al. [7] have proposed a pair-counting based measure named Normalized Probabilistic Rand (NPR) index to handle that case. In this article, we propose an information theoretic based measure, the Normalized Joint Mutual Information (NJMI), which is an extension of the Normalized Mutual Information (NMI).

In Sect. 2, we first review the information theoretic based measure NMI for comparing clusterings, and then describe the Joint Mutual Information (JMI) and its normalized version NJMI in detail. To validate the effectiveness of NJMI, the experimental results on some images selected from Berkeley segmentation database [1] are presented in Sect. 3. Section 4 gives the conclusion.

2 Normalized Joint Mutual Information for Segmentation Evaluation with Multiple Ground-Truth Images

In machine learning, Normalized Mutual Information is a widely used information criteria for clusterings comparison. It measures the similarity or distance between two clusterings by evaluating the mutual information between them. For image segmentation evaluation, we propose that this criteria is not only useful for binary segmentation evaluation, but it can also be generalized to the case of multiple ground-truth segmentations which is a hard problem and lacks good methodology to deal with.

2.1 Normalized Mutual Information for Binary Clusterings Comparison

Let D be a set of N data points $\{d_1, \ldots, d_N\}$, and U, V are two clusterings for D, where U includes r clusters $\{u_1, \ldots, u_r\}$, and V includes c clusters $\{v_1, \ldots, v_c\}$. If we regard U and V as two random variables of cluster labels, $p(u_i)$, $p(v_j)$ are the probabilities of a random data point labeled by u_i in U and labeled by v_j in V respectively, and $p(u_i, v_j)$ represents the joint probability that a data point labeled by u_i in U and v_j in V simultaneously, then according to information theory [8], the mutual information between random variables U and V is calculated as

$$I(U,V) = \sum_{i=1}^{r}\sum_{j=1}^{c} p(u_i, v_j) \log \frac{p(u_i, v_j)}{p(u_i)p(v_j)} \tag{1}$$

As the mutual information quantifies the information shared by U and V, it can also be used to measure the similarity between clustering U and clustering V. Further more, [4] proposed a normalized version of the mutual information which has fixed bounds $[0, 1]$:

$$NMI(U,V) = \frac{I(U,V)}{\sqrt{H(U)H(V)}} \tag{2}$$

where $H(U)$ and $H(V)$ are the entropies associated with U and V respectively, $H(U) = -\sum_{i=1}^{r} p(u_i)\log p(u_i)$, $H(V) = -\sum_{j=1}^{c} p(v_j)\log p(v_j)$.

2.2 Normalized Joint Mutual Information for Multiple Ground-Truth Segmentations

For the task of segmentation evaluation with multiple manually labeled images, we propose to use the Joint Mutual Information (JMI) to measure the similarity between the segmentation generated by an algorithm and a set of ground-truth images. Thus, for a given image I including N pixels, if set $U = \{U_1, \ldots, U_k\}$, abbreviated by $U_{1:k}$, denotes a set of ground-truth segmentations, variable V denotes a segmentation compared with U, the similarity measure is defined as

$$\begin{aligned} I\left(U_{1:k}; V\right) &= KL\left(p(u^{(1)}, \ldots, u^{(k)}, v) || p(u^{(1)}, \ldots, u^{(k)})p(v)\right) \\ &= \sum_{u^{(1)} \in U_1, \ldots, u^{(k)} \in U_k, v \in V} p(u^{(1)}, \ldots, u^{(k)}, v) \log \frac{p(u^{(1)}, \ldots, u^{(k)}, v)}{p(u^{(1)}, \ldots, u^{(k)})p(v)} \end{aligned} \tag{3}$$

where $KL(\cdot||\cdot)$ denotes the Kullback-Leibler divergence, $u^{(1)}, \ldots, u^{(k)}, v$ represent the variables about segment (class) labels in $U_1, \ldots, U_k, V$ respectively, and each of them may have different number of label values, e.g. $u^{(1)} = \{u_1^{(1)}, \ldots, u_n^{(1)}\}$ if there are n segments in U_1, and $u^{(2)} = \{u_1^{(2)}, \ldots, u_m^{(2)}\}$ if there are m segments in U_2. The joint probabilities $p(u^{(1)}, \ldots, u^{(k)}, v) = |u^{(1)} \cap \ldots \cap u^{(k)} \cap v|/N$ and $p(u^{(1)}, \ldots, u^{(k)}) = |u^{(1)} \cap \ldots \cap u^{(k)}|/N$ are the probabilities that a image pixel simultaneously assigned to segment labels $u^{(1)}, \ldots, u^{(k)}, v$ for $U_1, \ldots, U_k, V$, and simultaneously assigned to $u^{(1)}, \ldots, u^{(k)}$ for $U_1, \ldots, U_k$, respectively. Comparing Eq. (3) with Eq. (1), JMI can be seen as an extension of the mutual information.

Like mutual information, JMI does not have a fixed upper bound. To make the evaluation scores comparable in a fixed range $[0, 1]$, we also need a normalized version of JMI. In Sect. 2.1, the normalized mutual information has been given by

Eq. (2). In much the same way, we infer that JMI is bounded by the entropy $H(V)$ and the joint entropy $H(U_1, \ldots, U_k)$. So, Normalized Joint Mutual Information (NJMI) is defined as

$$NJMI = \frac{I(U_1, \ldots, U_k; V)}{\sqrt{H(U_1, \ldots, U_k)H(V)}} \tag{4}$$

where $H(U_1, \ldots, U_k) = -\sum_{u^{(1)} \in U_1, \ldots, u^{(k)} \in U_k} p(u^{(1)}, \ldots, u^{(k)}) \log p(u^{(1)}, \ldots, u^{(k)})$ and $H(V) = -\sum_{v \in V} p(v) \log p(v)$.

2.3 Joint Mutual Information and Multi-information

Further more, we illustrate that the JMI measure follows the intuitive principle that "'tightly knit' groups are more difficult to join" [9]. From the proof of Theorem 1 in [10], we obtain the relationship between the Joint Mutual Information and the Multi-information:

$$I(U_{1:k}; V) = \mathcal{I}(U_1, \ldots, U_k, V) - \mathcal{I}(U_1, \ldots, U_k) \tag{5}$$

where $\mathcal{I}(U_1, \ldots, U_k, V)$ and $\mathcal{I}(U_1, \ldots, U_k)$ are the two Multi-information quantities among multiple variables:

$$\mathcal{I}(U_1, \ldots, U_k, V) = \sum_{u^{(1)} \in U_1, \ldots, u^{(k)} \in U_k, v \in V} p(u^{(1)}, \ldots, u^{(k)}, v) \log \frac{p(u^{(1)}, \ldots, u^{(k)}, v)}{p(u^{(1)}) \ldots p(u^{(k)}) p(v)}$$

$$\mathcal{I}(U_1, \ldots, U_k) = \sum_{u^{(1)} \in U_1, \ldots, u^{(k)} \in U_k} p(u^{(1)}, \ldots, u^{(k)}) \log \frac{p(u^{(1)}, \ldots, u^{(k)})}{p(u^{(1)}) \ldots p(u^{(k)})}$$

In [9], the Multi-information is proposed to be used as a collective measure of similarity $s(i_1, i_2, \ldots, i_r)$ among $r > 2$ elements. So, if we put this relationship in the context of segmentation evaluation, supposing that U is a ground-truth set and V is a segmentation result, we observe that the value of JMI should be higher, if $\mathcal{I}(U_1, \ldots, U_k, V)$ is larger, i.e. the segmentation V and other ground-truth segmentations $\{U_1, \ldots, U_k\}$ are more similar to each other. Meanwhile, the JMI value is weakened by the collective similarity $\mathcal{I}(U_1, \ldots, U_k)$ among all of the segmentations in the ground-truth set. It indicates that if the manually segmented images are more consistent with each other, the compared segmentation needs to be more similar to them to get higher JMI value.

3 Experiment

In this section, we first present the performance of NJMI on comparing different segmentations of the same image. Figure 1 gives an example image and its four manually segmented images. Figure 2 shows seven mean shift segmentations (from oversegmentation to undersegmentation) using different bandwidth parameters. Figure 3 depicts three evaluation scores, Global Consistency Error

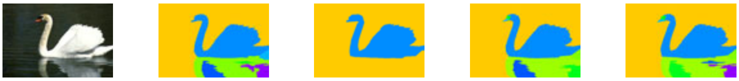
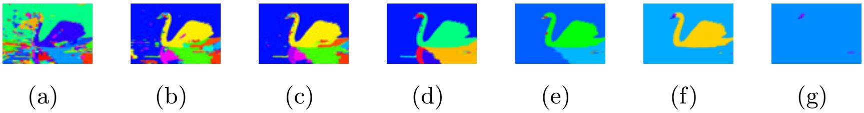


Fig. 1. An example image and its four ground-truth segmentations



(a) (b) (c) (d) (e) (f) (g)

Fig. 2. Seven mean shift segmentations

Table 1. Comparison of the average NJMI score of three segmentation algorithms on 50 images under the same number of segments

Number of Segments	mean shift	efficient graph	normalized cut
10	0.5829	0.5374	0.5923
20	0.5906	0.5695	0.6082
40	0.5903	0.5792	0.6124

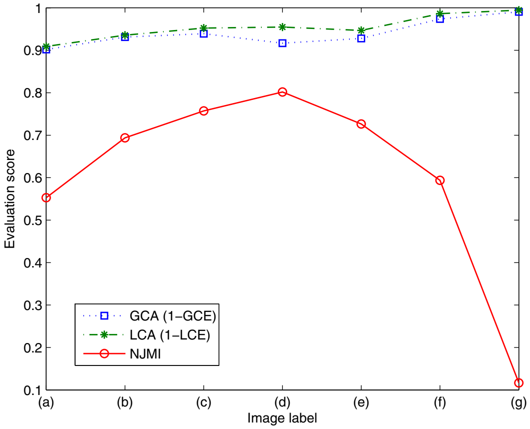


Fig. 3. Three evaluation scores for different segmentations: Global Consistency Accuracy (GCA), Local Consistency Accuracy (LCA), and NJMI

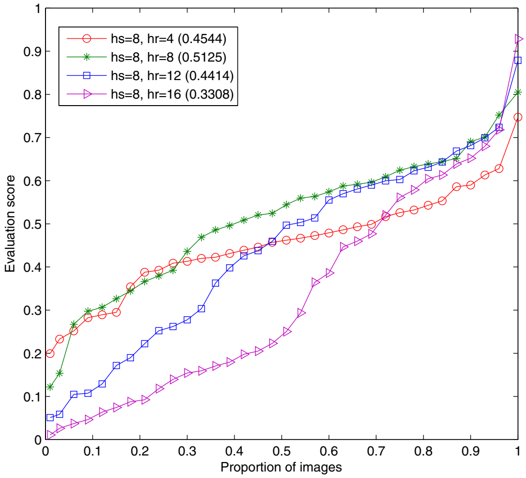


Fig. 4. Segmentation performance curves of mean shift algorithm. The number in parenthesis is the average NJMI score on all 100 images for the corresponding parameter setting (h_s is the scale bandwidth, h_r is the color bandwidth, and the minimum region $= 20$). The best parameter setting is $h_s = 8, h_r = 8$.

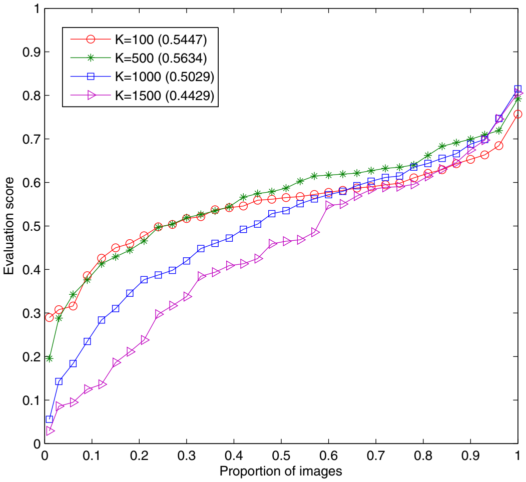


Fig. 5. Segmentation performance curves of efficient graph algorithm. The number in parenthesis is the average NJMI score on all 100 images for the corresponding parameter setting (parameter K controls the splitting process of a segment, and the minimum region $= 20$). The best performance is achieved when $K = 500$.

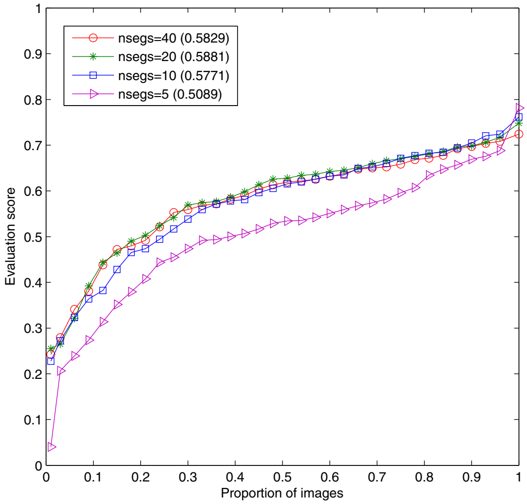


Fig. 6. Segmentation performance curves of normalized cut algorithm. The number in parenthesis is the average NJMI score on all 100 images for different target number of segments.

(GCE), Local Consistency Error (LCE) [1], and NJMI over the segmentations (a)-(g) in Fig. 2. From this plot, we observe that NJMI can indicate the segmentations with appropriate granularity being consistent with ground-truth images.

Furthermore, we explore the segmentation performance of three algorithms, mean shift (Fig. 4), efficient graph (Fig. 5) and normalized cut (Fig. 6) using NJMI cumulative-performance curve [11], which describes the performance distribution on 100 images selected from Berkeley segmentation database. In this curve, axis x represents the proportion of images, and axis y represents the NJMI score. A specific point $(x, f(x))$ on the curve indicates that $100 \cdot x$ percent of the images are segmented with a NJMI score lower than $f(x)$. Using a new segmentation method or a parameter setting produces a new curve, and the higher a curve, the better the performance of the corresponding parameter setting is. Therefore, we can roughly obtain the best parameter settings for mean shift and efficient graph. Then, we use these parameter settings for comparing the average performance of the three algorithms on another 50 images, under the same number of segments. The comparison results in Table 1 show that normalized cut has slightly better performance than the other two algorithms.

4 Conclusion

In this paper, we have presented an information theoretic based measure, the Normalized Joint Mutual Information, for image segmentation evaluation in the case of multiple ground-truth images being available. Experimental results show that NJMI can give reasonable scores which permit comparison between different segmentations of the same image. And it does not arbitrarily accommodate refinement or coarsening when all the human subjects give consistent segmentations for an image.

It should be noted that, to calculate NJMI, the joint probabilities need to be estimated first. This suffers from the curse of dimensionality when the ground-truth set is large. So, in that case, other parametric or non-parametric density estimation techniques should be brought in.

Acknowledgment. This work is supported by National Nature Science Foundation of China (60975078, 60902058, 60805041, 60872082), Beijing Natural Science Foundation (4092033, 4112047) and Doctoral Foundations of Ministry of Education of China (200800041049).

References

1. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In: ICCV (2001)
2. Rand, W.M.: Objective Criteria for the Evaluation Clustering Methods. Journal of the American Statistical Association 66(336), 846–850 (1971)
3. Meilă, M., Heckerman, D.: An Experimental Comparison of Model-based Clustering Methods. Machine Learning 42(1-2), 9–29 (2001)
4. Strehl, A., Ghosh, J.: Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions. J. Machine Learning Research 3, 583–617 (2002)
5. Meilă, M.: Comparing Clusterings by the Variation of Information. In: Conf. Learning Theory (2003)
6. Jiang, X., Marti, C., Irniger, C., Bunke, H.: Distance Measures for Image Segmentation Evaluation. EURASIP Journal on Applied Signal Processing, 1–10 (2006)
7. Unnikrishnan, R., Pantofaru, C., Hebert, M.: Toward Objective Evaluation of Image Segmentation Algorithms. IEEE Trans. Pattern Analysis and Machine Intelligence 29(6), 929–944 (2007)
8. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, Chichester (1991)
9. Slonim, N., Atwal, G.S., Tkačik, G., Bialek, W.: Information-based Clustering. PNAS 102(51), 18297–18302 (2005)
10. Zhou, Z., Li, N.: Multi-information Ensemble Diversity. In: El Gayar, N., Kittler, J., Roli, F. (eds.) MCS 2010. LNCS, vol. 5997, pp. 134–144. Springer, Heidelberg (2010)
11. Ge, F., Wang, S., Liu, T.: New Benchmark for Image Segmentation Evaluation. Journal of Electronic Imaging 16(3) (2007)

Alternating Scheme for Supervised Parameter Learning with Application to Image Segmentation

Lucas Franek and Xiaoyi Jiang

Department of Mathematics and Computer Science
University of Münster, Münster, Germany
{lucas.franek,xjiang}@uni-muenster.de

Abstract. This paper presents a novel alternating scheme for supervised parameter learning. While in previous methods parameters were optimized simultaneously, we propose to optimize parameters in an alternating way. In doing so the computational amount is reduced significantly. The method is applied to four image segmentation algorithms and compared with exhaustive search and a coarse-to-fine approach. The results show the efficiency of the proposed scheme.

1 Introduction

This work addresses the problem of supervised parameter learning for image segmentation algorithms. The developers of image segmentation algorithms typically train algorithm parameters in order to show how well an algorithm performs for a given dataset. Mostly, manual parameter training is not applicable for image segmentation. First, the algorithm has to be trained on a large database in order to get representative results, which causes considerable amount of manual work for a user. Secondly, if two different users tune parameters independently of each other by trial and error their results will be likely different. Therefore supervised parameter learning is often used in order to find the optimal parameter setting for a given database with ground truth [11].

In this paper, we propose a novel alternating scheme for supervised parameter learning. The idea behind our scheme is related to the principle used in the Expectation-Maximization algorithm: In each step all except for one parameter are hold fix and the best parameter setting for the free parameter is estimated. In the subsequent steps the procedure is repeated for each parameter until the improvement is small enough. The novel method is compared to some existing supervised parameter learning approaches: the exhaustive search and the coarse-to-fine approach [8]. The computational amount for this methods grows exponentially with the number of parameters. In contrast we will show that for the proposed alternating scheme the computational amount grows linearly with the number of parameters.

The rest of this paper is organized as follows. In Section 2 the alternating scheme for supervised parameter learning is proposed. Next, in Section 3 other

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 118–125, 2011.

existing supervised parameter learning methods used in our comparison are detailed. In Section 4 experimental settings are described and results are shown. Finally, we conclude in Section 5.

2 Alternating Scheme for Supervised Parameter Learning

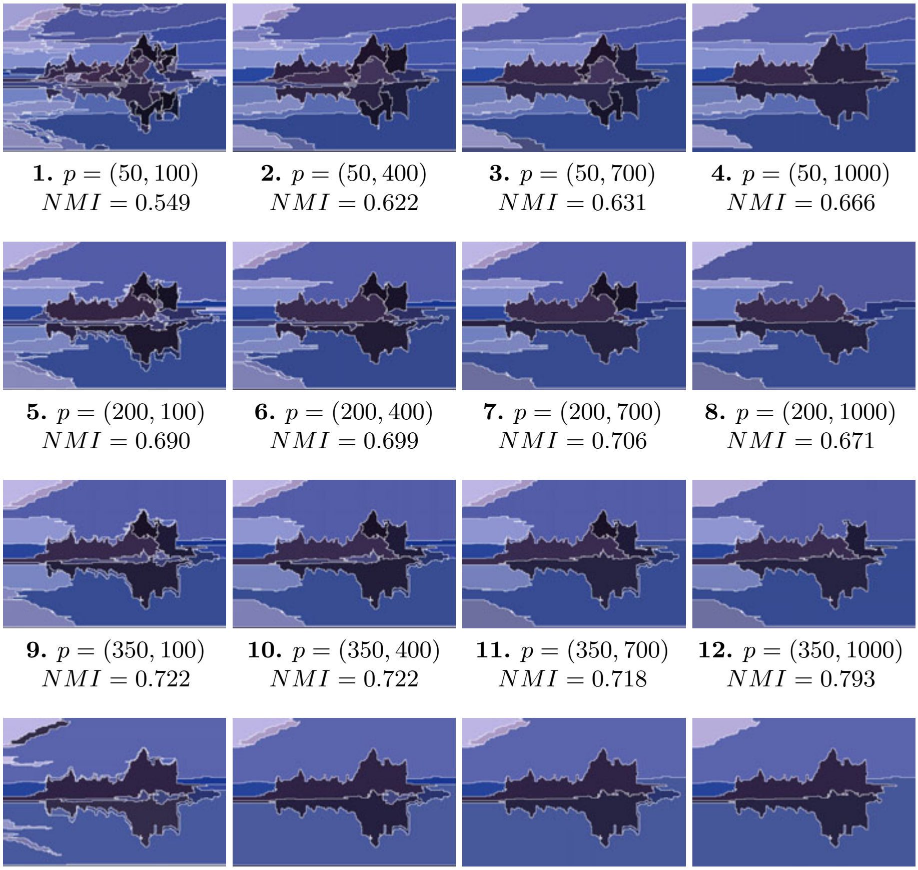
In this section we propose an alternating iterative scheme for supervised parameter learning. The exhaustive search and the coarse-to-fine approach [8] estimate all parameters simultaneously. In contrast, in the alternating iterative scheme all except for one parameter are fixed in each step and the best parameter setting for the free parameter is estimated. In the next step this parameter is fixed by using the estimated parameter setting and another parameter is set to be free. Within each iteration every parameter is set to be free once. By this way the computational complexity is reduced significantly.

Let us outline the case of two parameters in a more formal way:

1. Initialization: For each parameter an initial parameter range is sampled into N (assumed to be odd for notation simplicity) parameter settings.
2. For $i = 1$ to maximal number of iterations, do:
   (a) Fix one parameter by taking the median of the corresponding parameter range. Estimate the second parameter from the corresponding sampled parameter range. This is done by first segmenting each image of the given dataset N times by using the corresponding parameter settings. Then, the performance of the N segmentation results is computed and the parameter setting with the highest average performance is chosen.
   (b) Fix the second parameter trained from step (a) and estimate the first parameter from the corresponding sampled parameter range in the same way as it is done in step (a).
   (c) Reduce the parameter search ranges: The neighborhood of the selected parameter setting is resampled to obtain N parameter settings, $\frac{N-1}{2}$ to the left and $\frac{N-1}{2}$ to the right with a predefined step size. This step size should be appropriately chosen to narrow down the search space.
3. Output: The two parameter settings.

In our experiments the maximal number of 3 iterations yields good results. The scheme can be easily extended to an arbitrary number of parameters. This is simply done by alternatingly optimizing one of the parameters while fixing the others. The scheme is applicable to any segmentation algorithm with more than one parameter. Note that in case of only one parameter there is no alternation within each iteration and the scheme becomes similar to the coarse-to-fine approach [8]. For this alternating scheme the computational amount grows linearly with the number of parameters since each iteration DN segmentations (D: number of parameters) have to be computed.

We consider a simple example by varying the parameters of the segmentation algorithm FH [3]. FH has three parameters: a smoothing parameter (σ), a threshold function (k), and a minimum component size (min_size). For better illustration we fix in this case $\sigma = 0.6$ and vary the other parameters. The segmentation results and parameter settings are shown in Fig. 1. The image is taken from the Berkeley Dataset [6] and segmentations are evaluated by NMI [10]. Suppose the alternating scheme starts at the sixth segmentation. In this case the scheme proceeds as follows: In the first step $k = 200$ is fixed and min_size is estimated to be 700, i.e. the seventh segmentation is the best one in the second row. Now $min\_size = 700$ is fixed and $k = 500$ is estimated to be optimal in the third column (the 15th segmentation). The iteration stops because no further improvement is possible. Therefore $k = 500, min\_size = 700$ is estimated to be the best parameter setting for this single image.



1. $p = (50, 100)$ $NMI = 0.549$

2. $p = (50, 400)$ $NMI = 0.622$

3. $p = (50, 700)$ $NMI = 0.631$

4. $p = (50, 1000)$ $NMI = 0.666$

5. $p = (200, 100)$ $NMI = 0.690$

6. $p = (200, 400)$ $NMI = 0.699$

7. $p = (200, 700)$ $NMI = 0.706$

8. $p = (200, 1000)$ $NMI = 0.671$

9. $p = (350, 100)$ $NMI = 0.722$

10. $p = (350, 400)$ $NMI = 0.722$

11. $p = (350, 700)$ $NMI = 0.718$

12. $p = (350, 1000)$ $NMI = 0.793$

13. $p = (500, 100)$ $NMI = 0.785$

14. $p = (500, 400)$ $NMI = 0.832$

15. $p = (500, 700)$ $NMI = 0.832$

16. $p = (500, 1000)$ $NMI = 0.830$

Fig. 1. Segmentations generated by the FH algorithm and varying parameters $p = (k, min\_size)$, whereas the parameter $\sigma = 0.6$ is fixed. In each row (column) k (min_size) is constant. Segmentations are evaluated by NMI. Higher values are better.

3 Comparison of Supervised Parameter Learning Methods

In this section supervised parameter learning methods used in our comparison are detailed. The exhaustive search is the simplest method in order to learn the best parameter setting, but it is also the most time-consuming method. In this case the D-dimensional grid of parameter settings is sampled into N^D discrete parameter settings, resulting in a need of computing N^D segmentations, i.e. the computational amount grows exponentially with the number of parameters. In order to learn the parameters for a given dataset each image has to be segmented N^D times. Each segmentation has to be evaluated by a suitable performance measure (see Section 4) and the parameter setting candidate with the largest average performance is selected as the optimal parameter setting. Note that if the discretization of each parameter range is fine enough, the global optimum is approximated very well by the exhaustive search.

The coarse-to-fine approach proposed in [8] aims at reducing the computational effort by using a form of multi-locus hill climbing. The multi-dimensional grid of parameter settings in the first iteration is much smaller than for the exhaustive search. Several parameter settings yielding the best performance are estimated on this grid and further refined in the successive iterations. More specifically, the authors consider 5^D initial parameter settings in the first iteration. The highest performing one percent of the 5^D parameter settings are selected in order to refine the estimates in the second iteration. In the refinement step a 3^D sampling around each of the selected parameter settings is created. Again, the top-performing parameter settings are selected to be carried forward to the next iteration. The iteration continues until the improvement in performance is smaller than 1%. Also for this method the computational amount remains to grow exponentially with the number of parameters.

4 Experiments

In this section several experimental settings are detailed and then experimental results are shown.

4.1 Experimental Settings

First some experimental settings are summarised. In order to evaluates the similarity between the segmentation result and the given ground truth segmentation (GT) a similarity measure is needed. We decided to use two different similarity measures, the normalized mutual information (NMI) [10] and the boundary-based F-measure [7] to demonstrate the performance. Further, we use the Berkeley Database (BDS) with 300 images of size 321×214 (214×321 respectively). The BDS provides for each image several different GT segmentations. Therefore, we selected a representative GT segmentation, which maximizes the sum of similarity values (either NMI or F-measure) to the other ground truth segmentations. Because GT generation is a time-consuming task image databases with ground truth mostly contain much less than 300 images. In order to make our experiments more realistic, we decided to partition the BDS into 10 datasets, each consisting of 30 images. The datasets are numerated from 1 to 10. Note that while it is generally important to carefully design the partitioning of a database into training and test set [5], this issue is completely irrelevant in our study here.

We use four different segmentation algorithms to evaluate the proposed parameter learning method: the segmentation algorithm JSEG [2], the graph-based approach FH [3], the mean shift algorithm EDISON [1], and the Color Structure Code CSC [9]. The reasons for our choice are: 1) Most of these algorithms are state-of-the-art; 2) their code is available; 3) they are sensitive to parameter selection; 4) they have more than one parameter.

The JSEG algorithm has two parameters: $q \in [30, 600]$ and $m \in [0.05, 1]$. We have chosen to explore the following parameter ranges for FH: $\sigma \in [0.6, 1.5]$, $k \in [50, 500]$, $\mathit{min\_size} \in [100, 1000]$. The mean shift segmentation algorithm EDISON has three parameters: A feature (range) bandwidth $h_r \in [3, 21]$, a spatial bandwidth $h_s \in [3, 21]$, and a minimum region area (in pixels) $\mathit{min\_reg} \in [100, 1000]$. CSC is a hierarchical region growing method. In the HSV-mode three parameters: The hue-table $h \in [-4, 4]$, sat-table $s \in [-4, 4]$ and the val-table $v \in [-4, 4]$. For each channel it is possible to select one of nine distance tables containing thresholds for this channel. Further, we use for CSC a smoothing parameter $\sigma \in [0.6, 1.5]$ for Gaussian smoothing. Four parameters have to be optimized in this case.

Finally, let us mention some implementation details concerning the supervised parameter learning methods. In order to restrict the computational effort for the exhaustive search to a certain degree we chose for $D = 2$ an equidistant sampling of $N = 20$ samples for each parameter, while N is chosen to be 10 for $D = 3$. In case of CSC N is chosen to be 9 for parameters controlling the color thresholds and 10 for the smoothing parameter. For the coarse-to-fine approach we chose a subgrid of the grid used in the exhaustive search. Suppose the samples used in the exhaustive search are numerated from 1 to 10 (9, 20 respectively). For $D = 2$ and each parameter in the initial iteration of the coarse-to-fine approach the samples $2, 6, 10, 14, 18$ are used (these are 5 of the 20 samples used in the exhaustive search). Analogously, for $D = 3, 4$ the initial samples are chosen to be $1, 3, 5, 7, 9$. In the subsequent iterations the three top-performing samples are selected and refined by exploring the neighboring samples (in the grid computed for the exhaustive search).

For the alternating scheme N is chosen to be 9. For $D = 2$ the initial parameter settings of each parameter are set to the samples $2, 4, 6, 8, 10, 12, 14, 16, 18$ of the exhaustive search. The step size corresponds to the distance between two samples. For $D = 3, 4$ the initial samples of each parameter are set to be $1, 2, \ldots, 9$. In order to match always the samples on the grid of the exhaustive search the step size is set to the distance between two samples. By this way the comparison and computation get easier.

Table 1. JSEG algorithm. Comparison of supervised parameter learning methods: Exhaustive search (ExSearch), coarse-to-fine approach, and the proposed alternating scheme. Deviation is computed relative to the results of exhaustive search. For performance evaluation NMI and F-measure, respectively, are used. DS denotes the dataset.

	ExSearch	Coarse-to-fine		Alternating		ExSearch	Coarse-to-fine		Alternating	
DS	NMI	NMI	deviation in %	NMI	deviation in %	F	F	deviation in %	F	deviation in %
1	0.588	0.587	0.1	0.580	1.3	0.501	0.499	0.2	0.497	0.8
2	0.574	0.574	0.0	0.574	0.0	0.500	0.495	1.1	0.498	0.5
3	0.629	0.628	0.2	0.628	0.1	0.519	0.515	0.7	0.519	0.0
4	0.617	0.616	0.1	0.617	0.0	0.539	0.534	0.9	0.532	1.3
5	0.634	0.634	0.1	0.634	0.0	0.537	0.537	0.0	0.535	0.4
6	0.636	0.632	0.6	0.632	0.6	0.548	0.548	0.0	0.548	0.1
7	0.602	0.602	0.0	0.602	0.0	0.556	0.549	1.2	0.556	0.0
8	0.649	0.647	0.3	0.647	0.3	0.552	0.550	0.3	0.546	1.0
9	0.618	0.618	0.0	0.617	0.2	0.538	0.538	0.0	0.538	0.0
10	0.559	0.558	0.2	0.559	0.0	0.506	0.506	0.0	0.506	0.0

4.2 Results

In Table 1 the experimental results for the JSEG algorithm and each dataset are shown. For the coarse-to-fine approach and the alternating scheme the deviation (in %) from the results of the exhaustive search was computed. Note that the latter builds an upper bound of performance for the coarse-to-fine approach and the alternating scheme because a subgrid is used in these cases. The coarse-to-fine approach and the alternating scheme may converge to some local optimum. Therefore, their results are slightly worse than the exhaustive search. However, the deviation is always small and the found optima are acceptable in all cases. The results for FH, EDISON, and CSC are shown in Table 2, 3, 4, respectively. Also in these cases the deviation from the results of the exhaustive search is rather small, i.e. the optimum is approximated very well by the coarse-to-fine approach and the alternating scheme. Note that for CSC both the coarse-to-fine approach and the alternating scheme converge to the optimum even for all datasets. A close examination reveals that the main reason for this excellent convergence is that CSC is not as sensitive to its parameters as the other algorithms. Therefore, the optimum is easier to find in this case.

Table 2. FH algorithm. Comparison of supervised parameter learning methods.

	ExSearch	Coarse-to-fine		Alternating		ExSearch	Coarse-to-fine		Alternating	
DS	NMI	NMI	deviation in %	NMI	deviation in %	F	F	deviation in %	F	deviation in %
1	0.602	0.602	0.0	0.594	1.4	0.519	0.519	0.0	0.519	0.0
2	0.592	0.587	0.8	0.587	0.8	0.517	0.517	0.0	0.517	0.0
3	0.625	0.625	0.0	0.617	1.4	0.531	0.531	0.0	0.526	0.8
4	0.630	0.623	1.1	0.623	1.1	0.511	0.508	0.5	0.509	0.4
5	0.645	0.645	0.0	0.642	0.5	0.552	0.549	0.5	0.547	1.0
6	0.653	0.653	0.0	0.645	1.3	0.568	0.567	0.0	0.567	0.0
7	0.616	0.606	1.6	0.610	0.8	0.563	0.563	0.0	0.563	0.0
8	0.662	0.662	0.0	0.662	0.0	0.575	0.575	0.0	0.575	0.0
9	0.644	0.644	0.0	0.644	0.0	0.556	0.556	0.1	0.556	0.0
10	0.578	0.578	0.0	0.578	0.0	0.515	0.515	0.0	0.508	1.3

Table 3. EDISON algorithm. Comparison of supervised parameter learning methods.

	ExSearch	Coarse-to-fine		Alternating		ExSearch	Coarse-to-fine		Alternating	
DS	NMI	NMI	deviation in %	NMI	deviation in %	F	F	deviation in %	F	deviation in %
1	0.606	0.606	0.0	0.606	0.0	0.519	0.515	0.7	0.516	0.4
2	0.606	0.605	0.1	0.606	0.0	0.533	0.532	0.3	0.532	0.3
3	0.628	0.628	0.0	0.628	0.0	0.523	0.520	0.6	0.511	2.4
4	0.632	0.625	1.2	0.632	0.0	0.549	0.542	1.2	0.549	0.0
5	0.652	0.652	0.0	0.644	1.2	0.575	0.565	1.7	0.575	0.0
6	0.672	0.672	0.0	0.672	0.0	0.586	0.586	0.0	0.578	1.3
7	0.597	0.595	0.2	0.588	1.4	0.559	0.554	0.9	0.546	2.2
8	0.651	0.651	0.0	0.651	0.0	0.576	0.574	0.2	0.576	0.0
9	0.650	0.650	0.0	0.642	1.3	0.554	0.551	0.4	0.554	0.0
10	0.574	0.572	0.2	0.570	0.7	0.522	0.522	0.1	0.522	0.0

Table 4. CSC algorithm. Comparison of supervised parameter learning methods.

	ExSearch	Coarse-to-fine		Alternating		ExSearch	Coarse-to-fine		Alternating	
DS	NMI	NMI	deviation in %	NMI	deviation in %	F	F	deviation in %	F	deviation in %
1	0.499	0.499	0.0	0.499	0.0	0.454	0.454	0.0	0.454	0.0
2	0.514	0.514	0.0	0.514	0.0	0.429	0.429	0.0	0.429	0.0
3	0.565	0.565	0.0	0.565	0.0	0.459	0.459	0.0	0.459	0.0
4	0.527	0.527	0.0	0.527	0.0	0.441	0.441	0.0	0.441	0.0
5	0.567	0.567	0.0	0.567	0.0	0.477	0.477	0.0	0.477	0.0
6	0.586	0.586	0.0	0.586	0.0	0.496	0.496	0.0	0.496	0.0
7	0.522	0.522	0.0	0.522	0.0	0.484	0.484	0.0	0.484	0.0
8	0.614	0.614	0.0	0.614	0.0	0.523	0.523	0.0	0.523	0.0
9	0.533	0.533	0.0	0.533	0.0	0.452	0.455	0.0	0.455	0.0
10	0.490	0.490	0.0	0.490	0.0	0.467	0.467	0.0	0.467	0.0

Table 5. Runtime comparison of supervised parameter learning methods. Average runtime in minutes for a dataset with 30 images.

	D (# parameters)	ExSearch	Coarse-to-fine	Alternating
JSEG	2	465	41	41
FH	3	197	30	18
EDISON	3	8440	1245	760
CSC	4	4612	414	80

In Table 5 the runtime on a dual 2.66 GHz processor with 4 GB RAM is listed. In case of two parameters (JSEG) there is no clear difference between the coarse-to-fine approach and the alternating scheme. For FH the exhaustive search for a dataset with 30 images takes 197 minutes, while the computation time of the coarse-to-fine approach is 30 minutes. The alternating scheme yields a further speedup (18 minutes). For slower segmenters like EDISON the advantage of the alternating scheme is even more evident. The process needs 760 minutes for the alternating scheme, in contrast to 1245 and 8440 minutes for the coarse-to-fine approach and the exhaustive search, respectively. The largest speedup in comparison with the coarse-to-fine approach is observed for the case with four parameters (CSC): 80 minutes in contrast to 414 minutes for the coarse-to-fine approach. We conclude that in the case of three or more parameters the alternating scheme is clearly faster than the coarse-to-fine approach and therefore should be preferred in order to learn parameters.

5 Conclusion

Instead of estimating parameters simultaneously an alternating iterative scheme was proposed in this paper for supervised parameter learning. Experimental results in image segmentation have shown that the scheme converges quite well towards the global optimum. Further, it is much less time-consuming than the exhaustive search and the coarse-to-fine approach. Especially in case of a large number of parameters the speedup of the alternating scheme becomes significant.

The alternating scheme is applicable to other application domains, e.g. to range image segmentation or edge detection. [4] or edge detection. In future work we plan to extend the experiments to these domains.

References

1. Christoudias, C.M., Georgescu, B., Meer, P., Georgescu, C.M.: Synergism in low level vision. In: Proc. of Int. Conf. on Pattern Recognition, pp. 150–155 (2002)
2. Deng, Y., Manjunath, B.S.: Unsupervised segmentation of color-texture regions in images and video. IEEE Trans. Pattern Anal. Mach. Intell. 23, 800–810 (2001)
3. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. Int. J. Computer Vision 59, 167–181 (2004)
4. Jiang, X.: An adaptive contour closure algorithm and its experimental evaluation. IEEE Trans. Pattern Anal. Mach. Intell. 22(11), 1252–1265 (2000)
5. Jiang, X., Irniger, C., Bunke, H.: Training/test data partitioning for empirical performance evaluation. In: Christensen, H.I., Phillips, P.J. (eds.) Empirical Evaluation Methods in Computer Vision, pp. 23–37. World Scientific, Singapore (2002)
6. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. ICCV, vol. 2, pp. 416–423 (2001)
7. Martin, D., Fowlkes, C., Malik, J.: Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Trans. Pattern Anal. Mach. Intell. 26, 530–539 (2004)
8. Min, J., Powell, M., Bowyer, K.W.: Automated performance evaluation of range image segmentation algorithms. IEEE Trans. Systems Man and Cybernetics - Part B: Cybernetics 34, 263–271 (2004)
9. Rehrmann, V., Priese, L.: Fast and robust segmentation of natural color scenes. In: Proc. of Third Asian Conf. on Computer Vision, vol. 1, pp. 598–606 (1997)
10. Strehl, A., Ghosh, J.: Cluster ensembles - a knowledge reuse framework for combining multiple partitions. J. on Machine Learning Research 3, 583–617 (2002)
11. Unnikrishnan, R., Pantofaru, C., Hebert, M.: Toward objective evaluation of image segmentation algorithms. IEEE Trans. Pattern Anal. Mach. Intell. 29, 929–944 (2007)

Laser Line Segmentation with Dynamic Line Models

Jost Schnee and Jörg Futterlieb

Fraunhofer Institute for Factory Operation and Automation IFF,
Business Unit of Measurement and Testing Technology
Jost.Schnee@iff.fraunhofer.de

Abstract. Model based segmentation methods make use of a priori knowledge of the object to improve the segmentation results. Taking advantage of global information makes these methods less sensitive to local interferences like noise or line gaps. Laser lines from optical sensors are deformed due to the geometry of the measured object. Therefore we use dynamic models for robust and fast segmentation of a laser line.

1 Introduction

Using optical sensors based on laser triangulation for the acquisition of geometric data of objects has become a widespread method in the recent years. Nonetheless there are many tasks during the measurement process which still leave place for further improvements. In our work we focus on the task of segmenting the laser line in the image. This task is often regarded as a subtask of the peak detection. The methods in this article provide an advancement for the measurement process. They preliminarily determine the relative position of the laser line, so the peak detection can be realized at the laser line only. This makes the system robust to wrong measurements due to specular reflections of the laser light, or to intensity variations of the laser line. The segmentation is done with adapted versions of snakes [7] and the mass-spring model [2].

2 Detection of Laser Lines

Most research done in the field of image processing of laser line images is focused on the detection of the peak of the laser line. Fisher et al. [3] give a good comparison of most of the common peak detectors. They range from simple methods like linear interpolation or center of mass (sometimes called center of gravity) to more complex methods like Gaussian approximation, the parabolic estimator and the Blais and Rioux Detectors. A more sophisticated approach using FIR filters was introduced by Forest Collado in [4] and [5]. All peak detectors assume that the cross section of the laser line is nearly a Gaussian distribution superimposed by noise due to speckle. Thus they need the laser line to have a minimum width and intensity but not too many overexposed pixels. Depending on its complexity each peak detector has different requirements for the laser line.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 126–134, 2011.

The crucial point is that only little research is done to determine where to apply the peak detectors. The most common approach is to search for the maximum intensity peak in every row or column (depending on the alignment of laser and camera) in the image. Sometimes the first peak, which is higher than a certain threshold, is used to avoid secondary reflections when measuring translucent material. Another approach is to examine all peaks higher than a certain threshold and to remove erroneous data after the peak detection. This leads us to the segmentation of the laser line. Ofner et al. [8] proposed a line walking algorithm to extract line segments. It is looking for the maximum value in the adjacent row as long as it is higher than a certain threshold. The drawback of this method is that it segments a bulk of laser line segments instead of one line when it is confronted with line gaps. Line gaps appear when the intensity of the laser line is under the threshold in some areas. Such areas may occur, if the object consists of different materials or has changes on the surface due to corrosion or mechanical processing. To generate measuring data at these areas it is possible to use interpolation or approximation methods [10,8,9]. Nonetheless this is recommendable only when the measured object has a known geometry and the measurement task is to control a geometric feature like the radius of a circle. If the measurement task is surface inspection, it is better to apply local approximation methods for data smoothing only. This is because interpolation methods cannot distinguish between a gap in the line caused by lower intensities due to surface characteristics or a hole in the object.

3 Model Based Segmentation of Laser Lines

In our work we make use of the fact that we are searching for a laser line. Thus we describe the laser line as model and use this global information for segmentation. It allows laser line segmentation when there is no local information present at line gaps or the information is disturbed due to noise. The laser line from optical sensors is also deformed due to projection onto an object. Therefore we apply a smoothness constraint for controllable model deformation and use that dynamic line model for robust segmentation.

3.1 Mass-Spring Model

The first algorithm that we use for the segmentation of a laser line is based on the mass-spring model. It is an adaption of the physical concept of a mass-spring system. The structure and mode of operation of the model have been shown in detail in [2] and [1]. Thus we will give just a short aggregation of the

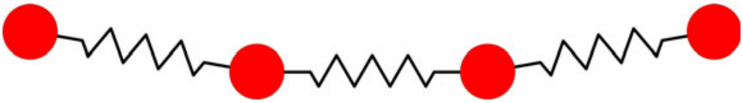


Fig. 1. Example line model (mass-spring model)

most important elements needed for our work. The system is composed of masses connected with elastic springs as shown in figure 1. The masses are representing image features and the springs are modeling their topology. Each mass is connected to sensors which generate external forces towards image features. Springs represent the internal energy and generate forces when they are compressed or stretched. During the segmentation process the internal and external forces will be balanced till a stable model state is reached. The force $\vec{F}_{spr_{ij}}$ of each spring is defined by the positions $\vec{p}$ of the connected masses i and j, the default length of the spring l_0 and the spring constant k_{ij} as follows.

$$\vec{F}_{spr_{ij}} = k_{ij}(||\vec{p_j} - \vec{p_i}|| - l_{0_{ij}})\frac{\vec{p_j} - \vec{p_i}}{||\vec{p_j} - \vec{p_i}||} \tag{1}$$

The external force $\vec{F}_{ext_i}$ affecting each mass, is calculated with an intensity sensor adapted to the segmentation of laser lines, which is introduced in subsection 3.3. The system is simulated by discrete time steps Δt. At each time step the velocity of the masses is calculated from the previous velocity $\vec{v_{i_t}}$ at t and the weighted forces at the masses, where w denote the weights and m the mass value. The damping constant d is used to suppress oscillations. To improve the stability of the model [1] introduced torsion forces $\vec{F}_{tor_{i,j}}$, which increase when the directions of the springs change related to their original direction. The weight of the torsion forces controls the allowed deformation of the line model.

$$\vec{v}_{i_{t+\Delta t}} = (\vec{v}_{i_t} + \frac{w_{spr}\sum_j \vec{F}_{spr_{ij}} + w_{tor}\sum_j \vec{F}_{tor_{ij}} + w_{ext}\vec{F}_{ext_i}}{m_i}\Delta t)(1-d) \tag{2}$$

The segmentation process is finished after fulfilling a custom criterion. In practice it has proved successful to cancel the computation when the sum of the movements of all masses is lower than a given threshold.

3.2 Snakes

The second algorithm that we use for the segmentation of a laser line is based on snakes, which have been published by [7] first. Snakes consist of an energy minimizing spline with control points. The spline builds the internal forces which impose a piecewise smoothness constraint. The external forces affect the control points and move them to the desired local minimum. In our application the external forces are derived from the intensities of the image. To characterize a line we changed the spline construction from a closed contour to an open contour as shown in figure 2. The energy function of the snake at the control points

Fig. 2. Example line model (snake model)

$p_i = (x_i, y_i)$ is shown in equation (3). E_{int} denotes the internal energy of the spline and E_{ext} the energy based on the image forces. The internal energy is the sum of a first order term and a second order term. The first order term specifies continuity (for each control point the two neighbored control points are taken into account) and is weighted with α_i. The second order term specifies curvature (for each control point the two previous and following control points are taken into account) and is weighted with β_i. The image energy E_{ext} is derived from the image intensities. Equation (4) shows the discrete notation of the energy function.

$$E^*_{snake} = \int\limits_0^1 E_{int}(i) + E_{ext}(i) \tag{3}$$

$$E^*_{snake} = \sum_{i=1}^{n} \alpha_i \left| p_i - p_{i-1} \right|^2 / 2h^2 + \beta_i \left| p_{i-1} - 2p_i + p_{i+1} \right|^2 / 2h^4 + E_{ext}(i) \tag{4}$$

To assign the line model to the image we need to minimize the energy and therefore we derive the internal energy. As mentioned in [7] the derivatives are approximated by finite differences, transformed into systems of equations and written as matrix as shown in equation (5).

$$Ax + f_x(x, y) = 0 \quad Ay + f_y(x, y) = 0 \tag{5}$$

The matrix A maps the internal energy on the control points and creates the internal forces. The number of rows and columns is equal to the number of points. We have constructed a matrix for a line (open contour) with $a_1 = \beta, a_2 = -(\alpha + 4\beta)$ and $a_3 = 2\alpha + 6\beta$ which looks as follows.

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ a_1 + a_2 & a_3 & a_1 + a_2 & 0 & 0 & \cdots & 0 & 0 & 0 \\ a_1 & a_2 & a_3 & a_2 & a_1 & 0 & \cdots & 0 & 0 \\ 0 & a_1 & a_2 & a_3 & a_2 & a_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & 0 & 0 & a_1 & a_2 & a_3 & a_2 & a_1 & 0 \\ 0 & \cdots & 0 & 0 & a_1 & a_2 & a_3 & a_2 & a_1 \\ 0 & 0 & \cdots & 0 & 0 & 0 & a_1 + a_2 & a_3 & a_1 + a_2 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{6}$$

The functions $f_x(x, y)$ and $f_y(x, y)$ denote the external forces on the control points. They are obtained by computing the intensity gradient at the position (x, y) of all control points with the intensity sensor introduced in subsection 3.3. The equations (5) are solved by the calculation of the time derivative at time t where γ denotes the step size for the time derivatives and I the identity matrix.

$$\begin{aligned} x_t &= (A + \gamma I)^{-1}(\gamma x_{t-1} - f_x(x_{t-1}, y_{t-1})) \\ y_t &= (A + \gamma I)^{-1}(\gamma y_{t-1} - f_y(x_{t-1}, y_{t-1})) \end{aligned} \tag{7}$$

The segmentation process assigns the line model to the image by calculating the equations (7) at each iteration. The process is finished after fulfilling a custom criterion. In practice it has proved successful to cancel the computation, when the sum of the movements of all control points is lower than a given threshold.

3.3 Intensity Sensor

To calculate the external forces $\vec{F}_{ext}$ we propose an intensity sensor which integrates the product of the intensities I_n of all pixels $\vec{p}_n$, in an defined circumcircle with the radius r around the position $\vec{s}$ of the sensor, and the vector towards them. Each vector is divided by its square length to degrade farther points. Integrating n pixels around the sensor suppresses speckle and noise.

$$\vec{F}_{ext}(\vec{s}) = \frac{1}{n} \sum_{n} \frac{(\vec{p}_n - \vec{s}) I_n}{|\vec{p}_n - \vec{s}|^2} \tag{8}$$

The radius of the intensity sensor is controlled during the segmentation process with equation (9). I_{max} denotes the maximum attainable intensity of the laser line, I_s the intensity at position $\vec{s}$, r_{max} the start radius and r_{min} the minimum radius.

$$r = r_{max} - (r_{max} - r_{min}) \min(1, \frac{I_s}{I_{max}}) \tag{9}$$

The automatic adaption of the radius of the intensity sensors enhances the segmentation speed and preserves the line ends since the radius determines the smoothing of the sensor at the line end.

3.4 Model Generation

Before the segmentation process with the line model can be started it is necessary to generate the model first. Thereto an area of interest around all pixels in the image with intensities greater than a given threshold is segmented. In this area a number of points representing the masses (mass-spring model) or control points (snake model) is created. To assure a good quality of the estimation of these starting points in relation to the laser line, the area of interest is divided into

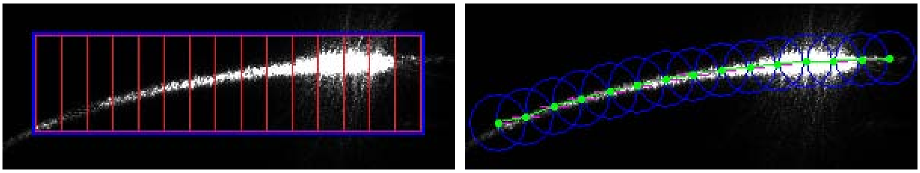

(a) (b)

Fig. 3. Generation of a line model for a horizontal laser line. The segments (red) of the area of interest (blue) for the line model (a), and the positions of the starting points (green) and the starting radii (blue) (b).

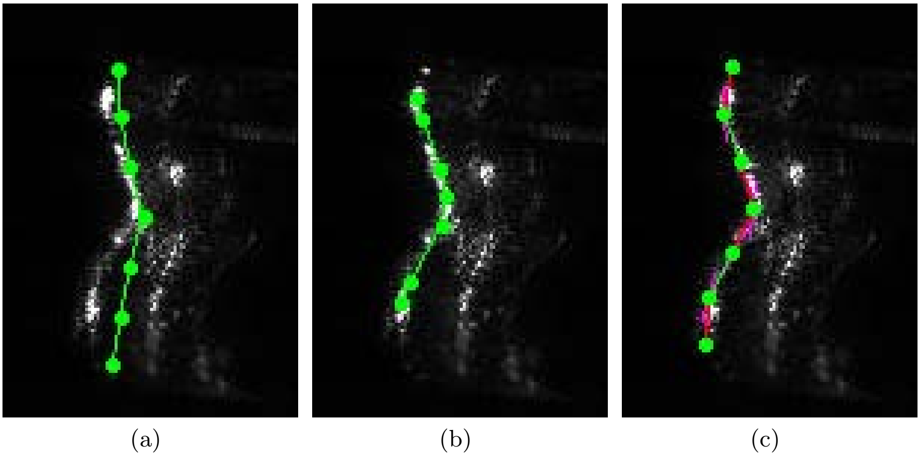


Fig. 4. Laser line on a transparent spectacle frame. The lines on the right side are secondary reflections from the back side of the spectacle frame. The starting position of both models (a), the final position of the snake model (b), and the final position of the mass-spring model (c).

segments [6]. Each starting point is positioned in the center of its segment along the preset orientation (horizontal/vertical) of the laser line as shown in figure 3. The offset of the starting points perpendicular to the preset direction of the laser line is calculated from the center of gravity of the intensity in these segments. In case of segments with very low intensities along the laser line, the positions of the starting points in these segments are calculated with linear interpolation between the next segments, which have starting points determined by an appropriate number of bright pixels.

4 Experimental Results

To examine the segmentation quality of the model based segmentation methods a set of images was used. These images have been chosen to represent a large variety of materials and applications. They contain images of specular reflections on metallic surfaces, diffuse reflections on rough surfaces, as well as specular reflections on transparent materials. The only constraint for the images is that the surface has to be continuous, so that there is only one line in each image.

4.1 Comparison of Mass-Spring Model and Snakes

The comparison of the segmentation quality of both methods shows almost similar results for all tested images. Depending on the laser line either the snake model or the mass-spring model provides a slightly better result. This can be seen in figure 4, where the starting position and the segmentation results on a spectacle frame are shown for both methods. The only difference is that the

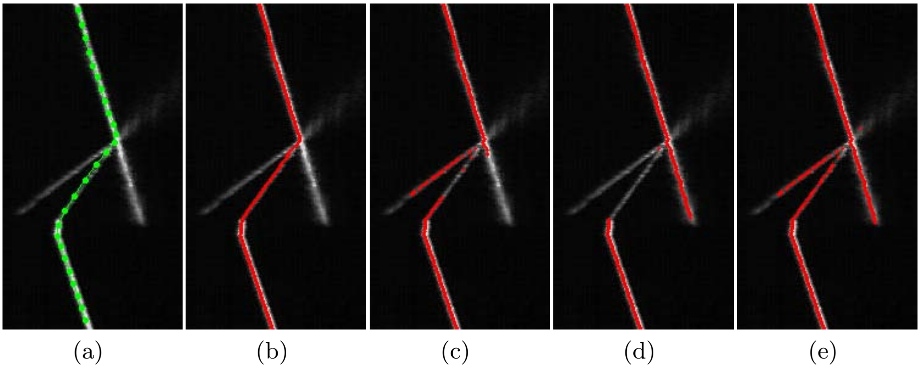


Fig. 5. Laser line with reflections in front of it and behind it. Segmentation result with snake model (a), peak detection with peak detector at snake positions (b), peak detection with first peak detector (c), peak detection with highest peak detector (d), and peak detection with all peaks detector (e). The detection starts from the left side with a threshold of 64[0,255] for all peak detectors.

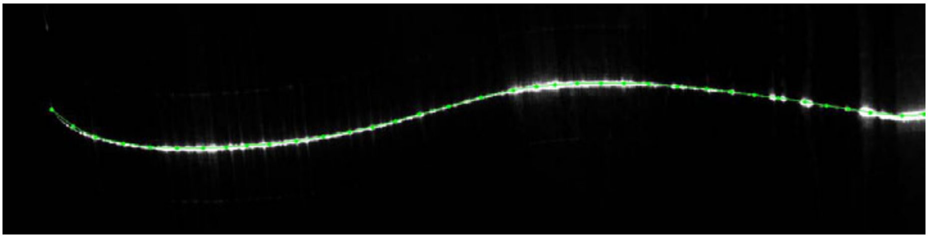


Fig. 6. Segmented laser line on a railway wheelset with high intensity variations due to mechanical wear and tear

springs of the mass-spring model hinder the compression and stretching of the model. Thus the line model is a bit longer there. Both methods are robust to the bad image quality due to speckle.

The implemented version of the snakes is faster than the implemented version of the mass-spring model. The computing time for the segmentation of a laser line in an image with the size of 1200x300 pixels, like the one shown in figure 6, is about 15 milliseconds for the snake model and 20 milliseconds for the mass-spring model (Intel Core Duo E6850 3GHz). Thus the method based on snakes is more suitable for the segmentation of a laser line than the method based on the mass-spring model.

Both methods have great advantages in addition to classic peak detection algorithms. They are able to segment a laser line even if it is disturbed by reflections. This is shown on the laser line in figure 5, which has specular reflections on metallic surfaces in front of it and behind it. The peak detection 5b along the segmented line model 5a determines the true position of the laser line. The peaks detected with the first peak detector 5c show some false detections on the reflection in front the laser line instead of on the laser line. The peaks detected with the highest peak detector 5d show some false detections on the reflection behind the laser line instead of on the laser line. And the detection of all peaks 5e shows false detections in front of the laser line and behind the laser line. Another advantage is that the methods are robust to high intensity variations along the laser line, as shown in figure 6.

5 Further Developments

The segmentation of the laser line before the peak detection steps makes it possible to use an automatic selection and parameterization of the peak detector accordant to the intensity of the laser line. This means that in bright parts more complex detectors which require a Gaussian distribution can be utilized while in dark parts simple detectors can be utilized. In addition the segmentation of the laser line gives the chance to evaluate the whole line with regard to the optimal exposure time of the camera. This evaluation should consider the continuity of the laser line without overexposed or underexposed areas. Besides it might be useful to apply an automatic split and merge algorithm [10] to the proposed methods. This will lead to better segmentation in case of multiple lines in the image.

6 Conclusion

The detection of the laser line is a fundamental step to determine the correct 3-D measurement data. Especially in industrial environments the optical imaging of the laser line is affected by reflections and has intensity variations due to different materials or to corrosion or mechanical processing. Thus it is a great advantage to segment the laser line before the peak detection step. The proposed methods have reduced the number of wrongly measured points caused by secondary reflections.

References

1. Dornheim, L.: Generierung und Dynamik physikalisch basierter 3D-Modelle zur Segmentierung des linken Ventrikels in SPECT-Daten. Diploma thesis. Otto-von-Guericke University of Magdeburg (2005)
2. Dornheim, L., Tonnies, K.D., Dornheim, J.: Stable dynamic 3D shape models (2005)
3. Fisher, R.B., Naidu, D.K.: A Comparison of Algorithms for Subpixel Peak Detection. In: Image Technology, pp. 385–404 (1996)
4. Forest, J., Salvi, J., Cabruja, E., Pous, C.: Laser stripe peak detector for 3d scanners. a fir filter approach. In: Proceedings of the 17th International Conference on Pattern Recognition, pp. 646–649 (2004)
5. Forest Collado, J.: New Methods for Triangulation-based Shape Acquisition using Laser Scanners. Ph.D. thesis, Universitat de Girona (2004)
6. Futterlieb, J.: Segmentierung von Laserlichtlinien. Term paper. Otto-von-Guericke University of Magdeburg (2011)

7. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: active contour models. International Journal of Computer Vision, 321–331 (1987)
8. Ofner, R., O'Leary, P., Leitner, M.: A collection of algorithms for the determination of construction points in the measurement of 3d geometries via light-sectioning, pp. 505–512 (1999)
9. O'Leary, P., Schalk, P., Ofner, R., Gfrerrer, A.: Instrumentation and analysis-methods for the measurement of profiles using light sectioning. In: Proceedings of IEEE Instrumentation and Measurement Technology Conference 2006, vols. 1-5, pp. 1108–1113 (2006)
10. Usamentiaga, R., Molleda, J., García, D.: Fast and robust laser stripe extraction for 3d reconstruction in industrial environments. Machine Vision and Applications, 1–18 (2010)

A Convex Active Contour Region-Based Model for Image Segmentation

Quang Tung Thieu[1], Marie Luong[1], Jean-Marie Rocchisani[2], and Emmanuel Viennet[1]

[1] L2TI, University Paris 13, Villetaneuse, France
{quangtung.thieu,marie.luong,emmanuel.viennet}@univ-paris13.fr
[2] Hopital Avicenne-Medecine Nucleaire, Bobigny, France
jean-marie.rocchisani@avc.aphp.fr

Abstract. A novel region-based active contour model is proposed in this paper. By using the image local information in the energy function, our model is able to efficiently segment images with intensity inhomogeneity. Moreover, the proposed model is convex. So, it is independent of the initial condition. Furthermore, the energy function of the proposed model is minimized in a computationally efficient way by using the Chambolle method.

Keywords: Segmentation, Convex, Medical Image, Active Contour.

1 Introduction

Image segmentation is one of the most important areas in image processing, with various applications such as medical imaging, where segmentation has been becoming a powerful computer-aided tool for cancer or pathological detection, diagnosis and treatment as well as for surgical planning [1,3,10]. Segmentation can provide measurements for the location, area, volume of desired object to detect, and information allowing a dynamical analysis of anatomical structures.

In the literature, one of numerous contributions gained the attention of scientists in the world, which concerns the original work of Kass *et al.* [4], introducing for the first time the active contour models (ACM), also called Snakes. Since then, Active Contour method has been developed and also proved to be one of the most robust methods for segmentation of medical images [1,2,3]. The key idea of these models relies on the use of curve evolution to detect objects in a given image. More precisely, the methods consist in deforming an initial curve towards object boundaries, under some constraints from that image. There are two main approaches for active contour models: the edge-based models [4,5,6,7,8] and the region-based models [9,10,11,17,18,19].

The edge-based models utilize the image gradient to guide the evolving curve toward object boundaries. Some methods, also referred as geometric active contours were firstly proposed by Caselles *et al.* [5] and Malladi *et al.* [6]. These models take benefice of the advantages of the level set method which is an implicit method and allows automatic change of topology. More recently, Shi *et al.*

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 135–143, 2011.

[16], proposed a fast method without the need of solving the partial differential equations. Some other methods concern the geodesic active contours models, also formulated using the level set [7,10]. However, as these edge-based models rely on the edge properties, their performance is reasonably satisfactory or even unsuccessful, especially in the case of weak boundaries. Other information can be useful to take into account, such as properties of regions between the contours.

The region-based models have been then proposed [9,11,13,17], which make use of statistical information of region instead of using image gradient, offering hence better performance in the case of noise and weak boundaries or discontinuous boundaries. One of the most well-known region-based models is the Chan-Vese model [11], which has been successful in handling images with homogeneous regions. However, by using global statistics, such method is not effective for segmenting objects with intensity inhomogeneity, as in the case of MR (Magnetic Resonance), PET (Positron Emission Tomography) or CT (Computed Tomography) images affected by shading artifact. The same authors have proposed the so-called Piecewise Constant (PC) models and extended their work for multiple regions using multiphase level set functions [17]. In [14], a PC convex model was proposed. Nevertheless, these models suffer the same drawback when treating images with inhomogeneous regions.

To cope with the problem of intensity inhomogeneity, several models have been proposed [17,18,19]. In particular, Chan and Vese also introduced in [17], the so-called Piecewise Smooth (PS) model. However, this model is complex and computationally expensive. More recently, Li *et al.* [18] proposed to use intensity information in local regions as constraints to deal with inhomogeneous regions. The proposed LBF (Local Binary Fitting) energy functional is based on the use of a kernel function to control the size of the neighbourhood. This problem is formulated and solved using the level set method. Later, Zhang *et al.* [19] proposed a Local Image Fitting (LIF) energy functional by minimizing the difference between an original image and the fitted image. Furthermore, a Gaussian kernel filtering is used to regularize the level set function after each iteration, avoiding hence re-initialization. Unfortunately, these models have the main drawback of non-convexity. So, the involved minimization problem may have local minima, which is not ideal for an automatic segmentation.

In this paper, we propose a convex active contour model to deal with intensity inhomogeneity. Due to the convexity properties, the proposed model is independent of the initial condition. In order to implement our model, we adopt the algorithm, which is introduced by Chambolle [20] for denoising and adapted by Bresson *et al.* in [14]. By using this algorithm, our proposed energy function is minimized efficiently in terms of computation. A comparison with other models such as the LBF and the LIF models has been performed to demonstrate the performance of our method. We first present a background in section 2. Then we introduce our proposed convex model and the implementation in section 3. Our results are presented in section 4 before the conclusion in section 5.

2 Background

Let Ω be the image domain, $u_0 : \Omega \to \mathbb{R}^+$ be a given image. Mumford and Shah [12] suggested to find a pair (u, C), minimizing the following energy function:

$$F^{MS}(u, C) = \int_\Omega |u - u_0|^2 dx + \rho \int_{\Omega \setminus C} |\nabla u|^2 dx + \mu |C| \qquad (1)$$

where u is a piecewise smooth approximation of u_0, while ρ, μ are positive constants, $|C|$ is the length of contour C. The segmentation is performed by minimizing this energy function. However, it is difficult to minimize this energy function directly because of the unknown set C.

To overcome this drawback, Chan and Vese [11] proposed an active contour model based on a simplified Mumford-Shah functional, with the assumption that solution image u is composed of two intensity piecewise constant regions. In this case, the second term of equation (1) is eliminated. Then, equation (1) becomes:

$$F^{CV}(c_1, c_2, C) = \int_{in(C)} |u_0 - c_1|^2 dx + \int_{out(C)} |u_0 - c_2|^2 dx + \mu |C| \qquad (2)$$

where μ is a positive constant, $out(C)$ and $in(C)$ represent the outside and inside regions of the contour C, respectively, c_1 and c_2 are two constants which approximate the image intensity in $in(C)$ and $out(C)$, respectively.

To minimize function (2), the contour C is replaced by the zero level set function [15]. However, the energy function is not convex. Then, there may be local minima and the resulting segmentation depends on the initial contour.

In order to eliminate the dependence on the initial contour, Chan *et al.* [13] proposed the following convex model:

$$\min_{c_1, c_2, 0 \le f \le 1} \int_\Omega |\nabla f| dx + \lambda \int_\Omega f(x)(c_1 - u_0)^2 dx + \lambda \int_\Omega (1 - f(x))(c_2 - u_0)^2 dx \quad (3)$$

where c_1, c_2 are the average intensities of $in(C)$ and $out(C)$, respectively. The resolution of this problem is performed by using the standard Euler-Lagrange equations technique and the explicit gradient descent based algorithm. However, the numerical scheme is very slow because of the regularization of the first term.

More recently, Bresson *et al.* proposed in [14] a similar model as follows:

$$\min_{c_1, c_2, 0 \le f \le 1} \int_\Omega g |\nabla f| dx + \lambda \int_\Omega f(x)(c_1 - u_0)^2 dx + \lambda \int_\Omega (1 - f(x))(c_2 - u_0)^2 dx \quad (4)$$

where g is an edge indicator function so that its value is small at object boundaries, e.g. $g(x) = \frac{1}{1 + \|\nabla u_0(x)\|}$. The authors also proved that if c_1 and c_2 are fixed, and if f^* is a minimizer of Eq. (4), then the set $M = \{x : f^*(x) > \alpha\}$ for any $\alpha \in (0, 1)$ determines a global minimizer of the Chan-Vese model.

Note that, when c_1 and c_2 are fixed, problem (4) becomes:

$$\min_{0 \le f \le 1} F(f) = \int_\Omega g(x) |\nabla f(x)| dx + \lambda \int_\Omega f(x) u_r(x, c_1, c_2) dx \qquad (5)$$

with $u_r(x, c_1, c_2) = (c_1 - u_0(x))^2 - (c_2 - u_0(x))^2$. Then, the constraint $0 \leq f \leq 1$ of (5) is eliminated according to the following claim [14].

Claim 1. *Let* $u_r(x, c_1, c_2) \in L^\infty(\Omega)$, for $c_1, c_2 \in \mathbb{R}$, $\lambda \in \mathbb{R}^+$, *then the convex constrained minimization problem (5) has the same set of minimizers as the following convex and unconstrained minimization problem:*

$$\min_f \int_\Omega g(x)|\nabla f(x)|dx + \int_\Omega (\lambda f(x) u_r(x, c_1, c_2) + \alpha\psi(f(x)))dx \tag{6}$$

where $\psi(z) = max\{0, 2|z - \frac{1}{2}| - 1\}$ *is an exact penalty function provided that the constant* α *is chosen large enough compared to* λ *such as* $\alpha > \frac{\lambda}{2}\|u_r(x)\|_{L^\infty(\Omega)}$.

Then, Bresson *et al.* [14] proposed a fast algorithm based on a dual formulation of the total variation norm, with the use of a fast algorithm proposed in [20] by Chambolle. Problem (6) is equivalent to the following problem:

$$\min_{f,v} \int_\Omega g(x)|\nabla f(x)|dx + \frac{1}{2\theta}\|f - v\|^2 + \int_\Omega (\lambda v(x) u_r(x, c_1, c_2) + \alpha\psi(v(x)))dx \tag{7}$$

where the parameter $\theta > 0$ is chosen to be small so that f and v are close.

The main drawback of these models is that they generally fail to segment images with inhomogeneous regions. Indeed, because c_1 and c_2 are the global average intensity inside and outside the contour, they can be far quite different from the original data if $in(C)$ or $out(C)$ are inhomogenous regions.

3 Our Model and Implementation

To segment images with inhomogeneous regions, we propose the following model which uses the edge information and local properties of regions inside and outside the evolved contour:

$$\min_{u_1, u_2, 0 \leq f \leq 1} \int_\Omega g(x)|\nabla f(x)|dx + \lambda \int_\Omega f(x) u_{in}(x)dx + \lambda \int_\Omega (1 - f(x)) u_{out}(x)dx \tag{8}$$

where g is defined in equation (4), λ is a positive constant, u_{in} and u_{out} are the data fidelity terms which are defined by the following equations:

$$\begin{aligned} u_{in}(x) &= \frac{\int_\Omega K_\sigma(x - y)(u_0(x) - u_1(y))^2 dy}{\int_\Omega K_\sigma(x - y)dy} \\ u_{out}(x) &= \frac{\int_\Omega K_\sigma(x - y)(u_0(x) - u_2(y))^2 dy}{\int_\Omega K_\sigma(x - y)dy} \end{aligned} \tag{9}$$

where K_σ is a Gaussian kernel with standard deviation σ, and $x, y \in \Omega$.

Here, we replace the global average intensities c_1 and c_2 in Eq. (3) or (4) by the approximation functions u_1 and u_2 of the local intensities inside and outside the contour C, respectively. It is easy to see that model (4) is a special case of our model if u_1 and u_2 are constants. The factor K_σ is added to express the localization property. Indeed, the Gaussian kernel $K_\sigma(x-y)$ should decreases rapidly to zero when y goes away from x. Thus, the energy $f(x)u_{in}(x)+(1-f(x))u_{out}(x)$ is dominated in the neighborhood of x.

Note that formula (9) is not similar to the formula of local intensity fitting [18] of the LBF model: First, the local energy function defined by $u_{in}(x)+u_{out}(x)$ is computed with the contribution of $u_1(y)$ and $u_2(y)$ evaluated for y varying over a neighborhood of a point x (integral of $u_1(y)$ and $u_2(y)$), while the local fitting energy of the LBF model is evaluated with the contribution of original $u_0(y)$ (integral of $u_0(y)$). Furthermore, we divide the weight function $K_\sigma(x-y)$ by $\int_\Omega K_\sigma(x-y)dy$, which can be considered as the area of the neighborhood of x. By this way, formula (8) is the general formula of (4).

To solve problem (8), implementation is performed by two steps as follows:

Step 1: Fixing the variable f, by using variation calculus method [21] for equation (8) with respect to u_1 and u_2, we obtain:

$$\begin{aligned} u_1(y) &= \frac{\int_\Omega K_\sigma(x-y)u_0(x)f(x)dx}{\int_\Omega K_\sigma(x-y)f(x)dx} \\ u_2(y) &= \frac{\int_\Omega K_\sigma(x-y)u_0(x)(1-f(x))dx}{\int_\Omega K_\sigma(x-y)(1-f(x))dx} \end{aligned} \tag{10}$$

Step 2: Fixing u_1 and u_2, the minimizer f^* of Eq.(8) is the same as:

$$\min_{0\le f\le 1} F(f) = \int_\Omega g(x)|\nabla f(x)|dx + \lambda\int_\Omega f(x)u_r(x,u_1,u_2)dx \tag{11}$$

with $u_r(x,u_1,u_2) = u_{in}(x) - u_{out}(x)$, $x\in\Omega$.

By Claim 1, the solution of (11) is the solution of the following problem:

$$\min_f \int_\Omega g(x)|\nabla f(x)|dx + \int_\Omega (\lambda f(x)u_r(x,u_1,u_2) + \alpha\psi(f(x)))dx$$

where $\psi(z)$ is defined as in (6). Then the variable v is introduced in:

$$\min_{f,v} \int_\Omega g|\nabla f|dx + \frac{1}{2\theta}\int_\Omega |f-v|^2dx + \int_\Omega (\lambda v(x)u_r(x,u_1,u_2) + \alpha\psi(v(x)))dx \tag{12}$$

where $\theta > 0$ must be chosen small sufficiently.

To solve the minimization problem (12), we use the fast algorithm based on a dual formulation of the total variation norm as proposed in [14,20]. After that, the set $M = \{x : f(x) > \alpha\}$ for any $\alpha\in(0,1)$ is used to extract the contours. The fast segmentation algorithm for solving (12) is resumed in Algorithm 1.

```
Algorithm 1
Input $u_0, g, \theta, \lambda, \tau, f$
  Repeat
    Calculate $u_1$, $u_2$ by Eq. (10)
    Calculate $u_{in}, u_{out}$ by Eq. (9)
    Calculate $u_r = u_{in} - u_{out}$.
    $v = \max\{\min\{f - \theta\lambda u_r, 1\}, 0\}$
    $p_0 := 0$
    Repeat
      $p^{n+1} = \frac{p^n + \tau\nabla\left(\mathrm{div}(p^n) - \frac{v}{\theta}\right)}{1 + \frac{\tau}{g}|\nabla\left(\mathrm{div}(p^n) - \frac{v}{\theta}\right)|}$
    To $p^{n+1} \approx p^n$
    $f := v - \lambda\mathrm{div}(p^{n+1})$
  To convergence
Output $f$
```

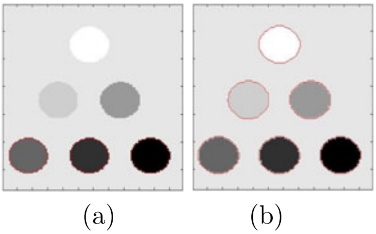

(a) (b)

Fig. 1. Segmentation results on synthetic image: (a) result of Chan-Vese convex model, (b) result of our model with $\theta = 0.01, \lambda = 1$

4 Experimental Results

To evaluate the performance of our method, several experimentations have been carried out on images with intensity inhomogeneity. Examples are shown here for some synthetic and MR as well as X-Ray images. A comparative evaluation has been performed to demonstrate the advantages of our method over other methods such as the LBF [18] and the LIF [19] models. The codes of the LBF and LIF models can be downloaded on the page of the author (http://www.engr.uconn.edu/∼cmli/ and http://www4.comp.polyu.edu.hk/∼cslzhang/).

We use the MATLAB r2008a to implement our algorithm. The program was run on a Dell (OptiPlex 360), which has Intel Core 2 Duo E7500 @ 2.93GHz and 4GB RAM. Here, we use the following values of parameters: $\tau = 0.01, \sigma = 3, \alpha = 0.5$. The other parameters are specified in figures.

First, the performance of our method for segmenting images of different distributions of intensity is evaluated. As shown in Fig. 1, our method succeeds in extracting all the objects of the synthetic image, including object with very similar intensity as the background. As the objects are piecewise constants, the Chan-Vese convex model [13] is also tested. As a result, this method only extracts three objects with intensities quite different from the background.

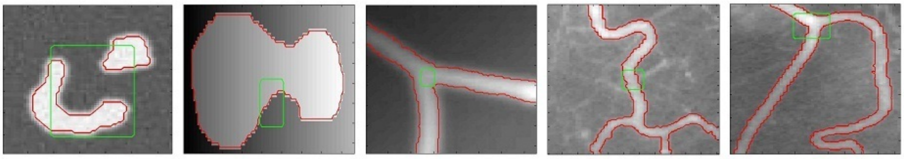


Fig. 2. Results of our model on images with intensity inhomogeneity. Green line: initial contour. Red line: final contour, $\theta = 0.01, \lambda = 1$.

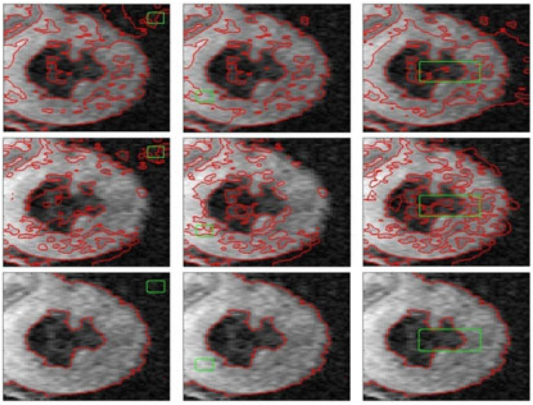


Fig. 3. Segmentation results of the LBF model (top row), LIF model (medium row) and our model (last row) on MR images. Green line: initial contour. Red line: final contour. $\theta = 0.1, \lambda = 0.1$.

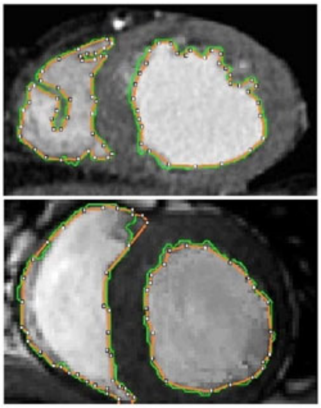


Fig. 4. Comparison with the ground truth established by expert on heart MR image. Yellow line: ground truth. Green line: our results.

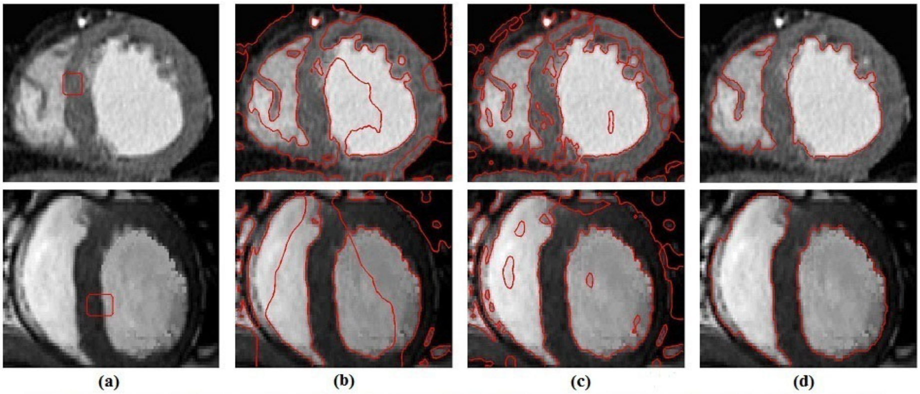


Fig. 5. Segmentation results on MR images: (a) initial counter; (b) results of the LBF model; (c) results of the LIF model; (d) results of our model. $\theta = 0.01, \lambda = 1$.

In order to demonstrate the performance of our method for addressing the intensity inhomogeneity, results obtained for typical images with non homogeneous regions such as blood vessel X-ray images and synthetic images are shown in Fig. 2. As can be seen, our method successfully achieves segmentation of objects of interest. The results of segmentation for MR images in Figs. 3, 4 and 5 show that our method gives accurate segmentation results, while the LBF and the LIF models fail to detect the truth contour. In particular, results of our method are compared with ground truth established by expert. In Fig. 4, an example is reported for the heart MR images. It is easy to see that the ventricle

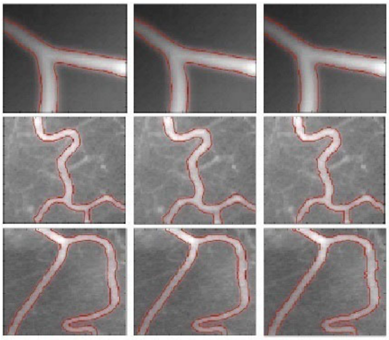
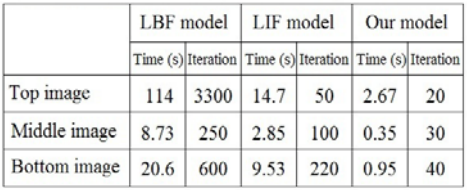


Fig. 6. The accurate segmentation results. From left to right: LBF, LIF and our models.

Table 1. Comparing CPU time (in second) and number of the iteration between our model with LBF, LIF models



	LBF model		LIF model		Our model	
	Time (s)	Iteration	Time (s)	Iteration	Time (s)	Iteration
Top image	114	3300	14.7	50	2.67	20
Middle image	8.73	250	2.85	100	0.35	30
Bottom image	20.6	600	9.53	220	0.95	40

boundaries of the heart are accurately extracted, as compared with the contour segmented by our expert. Another interest of our model is that it is convex. As can be seen in Fig. 3, our method gives satisfying results without depending on the initial contour, which is not the case for the LBF and LIF models.

Finally, we have made a comparative study for the CPU time using the same images and segmented results. As shown in Fig. 6 and Table 1, our method not only is faster but also takes less numbers of iterations than the other models.

5 Conclusion

In this paper, we proposed a novel region-based active contour model which is based on local information, allowing the model to segment images with intensity inhomogeneity. As the model is convex, the results obtained are independent of the initial contour. Furthermore, the implementation of our energy minimization model is performed using the dual formulation and the iterative algorithm of Chambolle. Experimental results have demonstrated the performance of our model in term of robustness to intensity inhomogeneity and computational time. Our model can be developped for 3D PET image segmentation and evaluated for its performance in terms of accuracy and computational time.

References

1. Xu, C., Pham, D.L., Prince, J.L.: Medical Image Segmentation Using Deformable Models. In: SPIE Handbook on Medical Imaging: Medical Image Analysis, pp. 29–174 (2000)
2. Boscolo, R., Brown, M., McNitt-Gray, M.: Medical Image Segmentation with Knowledge-guided Robust Active Contours. Radiographics 22(2), 437–448 (2002)

3. Wang, L., Li, C., Sun, Q., Xia, D., Kao, C.: Brain MR Image Segmentation Using Local and Global Intensity Fitting Active Contours/Surfaces. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part I. LNCS, vol. 5241, pp. 384–392. Springer, Heidelberg (2008)
4. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active Contour Models. Inter. J. Comp. Vi. 1, 321–332 (1988)
5. Caselles, V., Catte, F., Coll, T., Dibos, F.: A Geometric Model for Active contours in Image Processing. Numerische Mathematik 66, 1–33 (1993)
6. Malladi, R., Sethian, J.A., Vemuri, B.C.: A topology independent shape modeling scheme. In: Proc. SPIE Conf. Geo. Methods Comp. Vi. II, vol. 2031, pp. 246–258 (1993)
7. Caselles, V., Kimmel, R., Spiro, G.: On Geodesic Active Contours. The Inter. J. Comp. Vi. 22(1), 61–79 (1997)
8. Paragios, N., Deriche, R.: Geodesic Active Regions and Level Set Methods for Supervised Texture Segmentation. Inter. J. Comp. Vi. 46(3), 223 (2002)
9. Ronfard, R.: Region-based Strategies for Active Contour Models. Inter. J. Comp. Vi. 13(2), 229–251 (1994)
10. Chen, Y., Guo, W., Huang, F., Wilson, D.: Using Prior Shape and Points in Medical Image Segmentation. In: Rangarajan, A., Figueiredo, M.A.T., Zerubia, J. (eds.) EMMCVPR 2003. LNCS, vol. 2683, pp. 625–632. Springer, Heidelberg (2003)
11. Chan, T., Vese, L.: Active Contour without Edges. IEEE Trans. Ima. Process. 10(2), 266–277 (2001)
12. Mumford, D., Shah, J.: Optimal Approximations by Piecewise Smooth Function and Variation Problems. Comm. On Pure and App. Math. 42(5), 577–685 (1988)
13. Chan, T., Esedoglu, S., Nikolova, M.: Algorithms for Finding Global Minimizers of Image Segmentation and Denoising Models. J. App. Math. 66, 1632–1648 (2006)
14. Bresson, X., Esedoglu, S., Vandergheynst, P., Thiran, J., Osher, S.: Fast Global Minimization of The Active Contour/Snake Model. J. Math. Ima. Vi. 28(2), 151–167 (2007)
15. Osher, S., Sethian, J.A.: Front Propagating with curvature dependent speed: Algorithms Based on Hamilton-Jacobi Formulation. J. Com. Phy. 79, 12–49 (1988)
16. Shi,Y.,Karl, W.C.: A Fast Level Set Method Without Solving PDEs. In: ICASSP 2005 (2005)
17. Vese, L., Chan, T.: A Multiphase Level Set Framework for Image Segmentation Using The Mumford and Shah Model. Inter. J. Comp. Vi. 50(3), 271–293 (2002)
18. Li, C., Kao, C., Gore, J., Ding, Z.: Implicit Active Driven by Local Binary Fitting Energy. In: Proc. IEEE Conf. Comp. Vi. and Pat. Recog., pp. 1–7 (2007)
19. Zhang, K., Song, H., Zhang, L.: Active contours driven by local image fitting energy. Pat. Recog., 1199–1206 (2010)
20. Chambolle, A.: An Algorithm for Total Variation Minimization and Applications. J. Math. Imag. and Vi. 20(1-2), 89–97 (2004)
21. Aubert, G., Kornprobst, P.: Mathematical Problems in Image Processing: Partial Differential Equations and The Calculus of Variations. Springer, Heidelberg (2002)

Probabilistic Atlas Based Segmentation Using Affine Moment Descriptors and Graph-Cuts

Carlos Platero, Victor Rodrigo, Maria Carmen Tobar, Javier Sanguino, Olga Velasco, and José M. Poncela

Applied Bioengineering Group - Technical University of Madrid
carlos.platero@upm.es

Abstract. We show a procedure for constructing a probabilistic atlas based on affine moment descriptors. It uses a normalization procedure over the labeled atlas. The proposed linear registration is defined by closed-form expressions involving only geometric moments. This procedure applies both to atlas construction as atlas-based segmentation. We model the likelihood term for each voxel and each label using parametric or nonparametric distributions and the prior term is determined by applying the vote-rule. The probabilistic atlas is built with the variability of our linear registration. We have two segmentation strategy: a) it applies the proposed affine registration to bring the target image into the coordinate frame of the atlas or b) the probabilistic atlas is non-rigidly aligning with the target image, where the probabilistic atlas is previously aligned to the target image with our affine registration. Finally, we adopt a graph cut - Bayesian framework for implementing the atlas-based segmentation.

Keywords: probabilistic atlas, affine transformation, graph-cut.

1 Introduction

Much research has been developed to integrate prior knowledge into the segmentation task. We focus on prior knowledge of shape and appearance of the object of interest. These approaches requiere a modeling or training step before the actual segmentation takes place. These ideas find their root in the active shape models first introduced by Cootes et al [1] based on matching a shape model to an unseen image using landmarks. Later, there has been increasing interest in using level set-based representation for shape priors [2,3,4], which avoids landmarks. However, segmentation techniques that rely on the optimization of the complex functionals require the adjustment of multiple parameters. Consequently, these methods suffer from sensitivity to the tuning process.

In medical images there is sometimes a weak relation between voxel data and the label assignment. In such cases, spatial information must be taken into account in the segmentation process. One well-validated approach relies on combining the segmentations obtained from non-rigid aligning multiple manually labeled atlas with the target image [5]. This method makes no use of the intensity information. Considering such information could improves the quality of

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 144–151, 2011.

the atlas segmentation. The probabilistic atlas is commonly used in the analysis of medical images, since it integrates a priori knowledge of the shape and the appearance.

To combine prior knowledge and some type of regularization based on the framework of a Markov random field (MRF) is a well established technique for the medical image segmentation [6]. In these approaches spatial information in terms of a probabilistic atlas and the contextual information are used to formulate a maximum a posteriori probability (MAP-MRF). Since Grey et al [7] proposed graph cuts as a generic method for estimating the maximum a posteriori, it has been widely used for optimization in this area. For the case of two labels, Greig constructed a graph with two terminal vertices such that the minimum cut provides a global optimal labeling. For the multi-label problem, Ishikawa [8] solved the minimization problem for energy functions with pairwise terms that are convex in the linearly ordered labels. Therefore, we adopt a graph cut - Bayesian framework for implementing the atlas-based segmentation.

The paper is organized as follows: in Section 2, we show the problem of the linear registration and our approach based on the image normalization. Section 3 describes the method for constructing a probabilistic atlas. Section 4 presents our framework for segmentation using the atlas information and the graph cuts for optimizing the posterior probability. Finally, in Section 5, we apply our procedure to liver segmentation from CT images.

2 Linear Registration Using Affine Moment Descriptors

We present a procedure for generating a probabilistic atlas based on affine moment descriptors. It captures the variability of learning samples and tries to generalize for the segmentation task. Our first step is to align the training samples, in order to avoid artifacts due to different pose. Traditionally, the pose parameters have been estimated minimizing a energy functional, via gradient descent [3,9].

Our approach considers a training set consisting of N binary images $\{S_i\}_{i=1,\dots,N} : \Omega \subset \mathbb{R}^n \rightarrow \{0,1\}$, $n = 2$ or 3. For the multi-label problem, manual segmentation images have to be converted to binary images. All images in the database are aligned with a single binary image as reference, $\tilde{S}_{ref}$. The new aligned images are defined as $\tilde{S}_i = S_i \circ T_i^{-1}$, where T_i is an affine transformation, given by the composition of a rotation, a scaling transformation and a translation. Equivalently, $S_i = \tilde{S}_i \circ T_i$. We propose a criterion for alignment based on a shape normalization algorithm [11]. It is only necessary to compute the first and second order moments. The first-order moments locate the centroid of the shape and the second-order moments characterize the size and orientation of the image. Given a binary image S_i, we compute the second-order moments matrix, and the image is rotated using the eigenvectors and it is scaled along the eigenvectors according to the eigenvalues of the second-order moment matrix of S_i and $\tilde{S}_{ref}$, where $\tilde{S}_{ref}$ is a normalized shape. Then, it is translated by the centroid. We do not consider the problem of reflection (for this see [10]). If we only consider moments up to order two, S_i is approximated to an ellipse/ ellipsoid centered at the image centroid. The rotate angles and the axes are determined by the eigenvalues and the eigenvectors of the second-order moment matrix [11]. Let R_i be the rotation matrix.

Let $\{\lambda_j^{ref}\}_{j=1,..,n}$ be the eigenvalues of the reference image $\tilde{S}_{ref}$. We consider one of the following scale matrices: a) $W_i = \sqrt{\frac{\lambda^{ref}}{\lambda^i}} \cdot I$ where $\lambda^c = (\prod_{j=1}^n \lambda_j)^{1/n}$, $c = \{ref, i\}$ and I is the identity matrix or b) W_i is a diagonal matrix where $w_{j,j} = \sqrt{\frac{\lambda_j^{ref}}{\lambda_j^i}}$. In the first case it is a homothety, while in the second case the size fits in each principal axis as the reference. The first option is used for shape priors without privileged directions otherwise the second case is chosen. Finally, the affine transformation translates the origin of the coordinate system to the reference centroid $\overline{x}_{ref}$. We denote the $i-$shape centroid as $\overline{x}_i$. The affine transformation is then defined as follows:

$$T_i^{-1}(x) = R_i \cdot W_i \cdot (x - \overline{x}_i) + \overline{x}_{ref}. \tag{1}$$

This affine transformation aligns from S_i to $\tilde{S}_{ref}$. Of course, it is a bijection if $det(R_i \cdot W_i) \neq 0$. If we use a scaling identical in all directions, $\tilde{S}_{ref}$ will be only a numeric artifact for the pose algorithm. The alignment error does not depend on the reference, $\tilde{S}_{ref}$. But when each principal axis is adjusted to the reference, the alignment error depends on the choice of the reference. We can not guarantee the optimal pose for any shape. But neither the gradient descent method guaranteed to find the optimum because there is not evidence that the proposed functionals are convex. Our procedure is fast and optimum if the shapes are closed to ellipses or ellipsoids.

3 Construction of the Probabilistic Atlas

Our framework is based on the Bayesian decision theory. Given the target image to be segmented, $I : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}$, and the probabilistic atlas, it assigns the label that maximizes the posterior probability:

$$F_x = F(x) = \arg\max_{l_j \in L} p(I_x|l_j)p(x, l_j),$$

where $F : \Omega \subset \mathbb{R}^n \rightarrow L = \{l_1, l_2, \ldots, l_k\}$ is a labeling of the voxels of Ω, $p(I_x|l_j)$ represents the likelihood term of the voxel appearance at x corresponding to the label l_j and $p(x, l_j)$ is the prior term at x, which models the shape variability. Therefore, these terms represent the appearance and shape models and they are constructed using the aligned training images. The appearance and shape models are built with the variability of our linear registration.

3.1 Appearance Prior Modeling

The appearance model is obtained from the intensities of the voxels belonging to the set of aligned training images. Before building the appearance model, the training images in intensity are normalized using histogram matching. We denote the normalized and aligned training images as $\{\tilde{S}_i\}_{i=1,..N} : \tilde{\Omega} \subset \mathbb{R}^n \to L$ and $\{\tilde{I}_i\}_{i=1,..N} : \tilde{\Omega} \subset \mathbb{R}^n \to \mathbb{R}$. Gaussian mixture models are used intensively for distribution estimation and their parameters are tuned by using expectation-maximization based method [6], which provides a global view of the whole object appearance. In this work we model the probabilistic appearance for each voxel and each label using parametric or nonparametric distributions. We have implemented two options: i) each voxel on the aligned learning set follows a normal distribution for each label, $N(\mu(x,l_j), \sigma_G^2(x,l_j))$:

$$\mu(x,l_j) = \frac{\sum_{\{i|\tilde{S}_i(x)=l_j\}} \tilde{I}_i(x)}{\#\{i|\tilde{S}_i(x)=l_j\}} \qquad \sigma_G^2(x,l_j) = \frac{\sum_{\{i|\tilde{S}_i(x)=l_j\}} (\tilde{I}_i(x) - \mu(x,l_j))^2}{\#\{i|\tilde{S}_i(x)=l_j\}}.$$

ii) It follows a nonparametric distribution, considering the probabilistic atlas and the target image into the same coordinate frame (see next section):

$$p(\tilde{I}_x|l_j) = \frac{1}{\#\{i|\tilde{S}_i(x)=l_j\}\sigma_W(x,l_j)} \sum_{\{i|\tilde{S}_i(x)=l_j\}} K\left(\frac{\tilde{I}(x) - \tilde{I}_i(x)}{\sigma_W(x,l_j)}\right),$$

where $K(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{|z|^2}{2})$ and
$\sigma_W^2(x,l_j) = \frac{1}{\#\{i|\tilde{S}_i(x)=l_j\}} \sum_{\{p|\tilde{S}_p(x)=l_j\}} \min_{p \neq q}(\tilde{I}_p(x) - \tilde{I}_q(x))^2$.

3.2 Shape Prior Modeling

To capture the variability of the shape, the set of aligned manually segmented images are used. In [2,3,4], principal component analysis (PCA) of the signed distance functions of training data is used to capture it. However, PCA provides a global view of the shape variability. Saad et al [12] introduce a modification of the idea of a probabilistic atlas by incorporating additional information derived from the distance transform. However, we have observed that a local estimation over our aligned training data provides robust results. It defines the prior term at x applying the vote-rule as

$$\sum_{j=1}^{k} p(x,l_j) = 1 \qquad p(x,l_j) = \frac{\#\{i|\tilde{S}_i(x)=l_j\}}{N}.$$

4 Image Segmentation Strategy

The appearance and shape models are built with the variability of our linear registration, without learning based on non-rigid registration as in [5]. The drawback is the need of an initial solution for the segmentation. However, this initial solution does not need to be robust because the proposed affine transformation uses only the first and second order moments.

Given a new image to be segmented, I, and an initial binary solution, $S : \Omega \subset \mathbb{R}^n \to \{0,1\}$, we have two procedures: a) we apply the proposed affine registration to bring this image into the coordinate frame of the atlas $\tilde{S} = S \circ T^{-1}$ and $\tilde{I} = I \circ T^{-1}$, where T^{-1} is calculated as (1) or b) the probabilistic atlas is non-rigidly aligning with the target image, where the probabilistic atlas is previously aligned to the target image with our affine registration, $S_{ref} = \tilde{S}_{ref} \circ T$. In both cases, the posterior probability is calculated for the unseen image.

Optimizing the posterior probability is not an easy task, especially because there are so many realizations of the MRF model and the optimization is prone to be caught in local maximums. Greig et al. [7] were the first to discover that powerful min-cut/max-flow algorithms from combinatorial optimization can be used to minimize certain important energy functions in computer vision. In particular, they showed that graph cuts can be used for restoration binary images. The problem was formulated as a maximum a posterior estimation with a MRF regularization that required the minimization of the following energy:

$$E(F) = \sum_{x \in \Omega} V_x(F_x) + \sum_{\substack{\{x,y\} \\ x,y \in \Omega, x \neq y}} V_{xy}(F_x, F_y), \tag{2}$$

where for the case of two labels $L = \{l_1, l_2\}$, $V_x(F_x) = \begin{cases} \lambda_x & \text{if } F_x = l_2 \\ 0 & \text{if } F_x = l_1 \end{cases}$, $\lambda_x = \log\left(\frac{p(I_x|l_2)p(x,l_2)}{p(I_x|l_1)p(x,l_1)}\right)$ and $V_{xy}(F_x, F_y) = \begin{cases} \beta_{xy} > 0 & \text{if } F_x \neq F_y \\ 0 & \text{if } F_x = F_y \end{cases}$. Greig constructed a graph with two terminal vertices $\{s, t\}$, such that the minimum cut provides a global optimal labeling. There is a directed edge $\{s, x\}$ from s to the voxel x with weight $\omega_{sx} = \lambda_x$ if $\lambda_x > 0$; otherwise, there is a directed edge $\{x, t\}$ from x to t with weight $\omega_{xt} = -\lambda_x$. There is an undirected edge $\{x, y\}$ between two internal vertices with weight $\omega_{xy} = \beta_{xy}$. It is a smoothness term based on intensities $\{I(x), I(y)\}$, which represents the realizations of the MRF model. For the multi-label problem, if each V_{xy} defines a metric, then minimizing (2) is known as the metric labeling problem and can be optimized effectively with the α-expansion algorithm [13].

5 Validation, Experiments and Results

The experimental validation is performed using the problem of liver segmentation from 3D CT images. Algorithms relying solely on image intensities or derived features usually fail. To deal with missing or ambiguous low-level information, shape and appearance prior information has to be employed. The proposed method has been considered on 20 patients CT slice set and tested on another 10 specified CT datasets.

In a first step we align the training data by the proposed procedure. In this case, each principal axis is adjusted to the reference. Experimentally, $\tilde{S}_{ref}$ was chosen by minimizing the *Similarity Index*, $\left(SI = \frac{1}{N} \sum_i \frac{(S_i \circ T_i^{-1}) \cap \tilde{S}_{ref}}{(S_i \circ T_i^{-1}) \cup \tilde{S}_{ref}}\right)$, over the training set. We compare our approach with other techniques. Table 1 lists the mean (μ_{SI}) and standard deviation (σ_{SI}) values of the SI metric over the training data set. Given a CT abdominal image as target image, our approach starts

Table 1. Results of the affine registration: μ_{SI} and σ_{SI} values of the SI

Type	SSD [3]	MI [9]	Our
μ_{SI}	0.54	0.63	0.68
σ_{SI}	0.10	0.07	0.05

with an initial solution. It is obtained filtering the image by a nonlinear diffusion filter with selection of the optimal stopping time. Then, region growing and 3D edge detector are applied to the filtered image. Morphological post-processing merges the previous steps, giving the initial solution. Next, the probabilistic atlas and the target image are placed into the same coordinate frame using our affine transformation. The non-rigid registration between the probabilistic atlas and the target image was performed with ElastiX [14]. We use the min-cut/max flow algorithm of Boykov-Kolmogorov for energy minimization [15]. In our implementation, the data term, defines the edge weights connecting each node to the source s and sink t, is proposed to: $\lambda_x = \log\left(\frac{\frac{p(I_x|l_2)p(x,l_2)}{\sigma(x,l_2)}}{\frac{p(I_x|l_1)p(x,l_1)}{\sigma(x,l_1)}}\right)$. This proposal is based on the more reliable of the probability estimations if there are less dispersion in the samples. We have experimentally observed that normal distribution model for probabilistic appearance prior is more robust than the non-parametric one. We think that it is due to less dependence on the initial solution. The parameters of V_{xy} were tuned using the leave-one-out technique from training data according to the segmentation scores. In our case, a 6 neighborhood relation is used to save memory.

Fig. 1 shows slices from two cases, drawing the result of the method (in blue) and the reference (in red). The quality of the segmentation and its scores are based on the five metrics [16]. Each metric was converted to a score where 0 is the minimum and 100 is the maximum. Using this scoring system one can loosely say that 75 points for a liver segmentation is comparable to human performance.

Table 2. Average values of the metrics and scores for all ten test case: volumetric overlap error (m_1), relative absolute volume difference (m_2), average symmetric surface distance (m_3), root mean square symmetric surface distance (m_4) and maximum symmetric surface distance (m_5)

Type		m_1	m_2	m_3	m_4	m_5
AT1 [5]	metrics	12.5%	3.5%	2.41 mm	4.40 mm	32.4 mm
	scores	51	80	40	40	57
Affine	metrics	12.1%	2.5%	1.71 mm	2.96 mm	26.1 mm
	scores	53	87	57	59	66
Nonrigid	metrics	9.69%	3.9%	1.12 mm	2.03 mm	22.1 mm
	scores	62	79	72	72	71

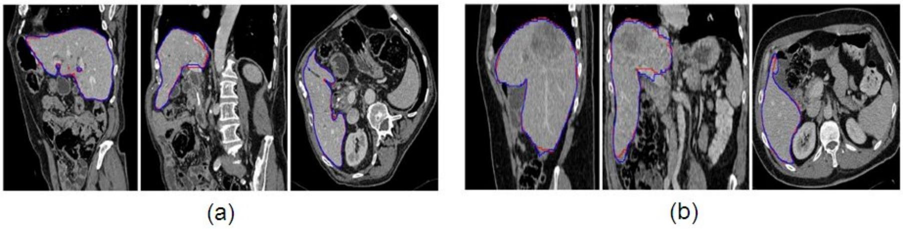


Fig. 1. From left to right, a sagittal, coronal and transversal slice for an easy case (a) and a difficult one (b). The outline of the reference standard segmentation is in red, the outline of the segmentation of the method described in this paper is in blue.

Table 2 lists the average values of the metrics and their scores over the test data set. It shows the performances for the three segmentation strategies: a) atlas matching b) probabilistic atlas with only linear registration and b) probabilistic atlas with nonrigid techniques. The average computation times for the liver segmentation task are 203.4 s., 25.3 s. and 211.7 s. respectively ([Dual CPU] Intel Xeon E5520 @ 2.27GHz).

6 Conclusion

We have presented two main contributions. Firstly, the linear registration has been solved using an image normalization procedure applied to the labeled atlas. An advantage is that the proposed affine transformation is defined by closed-form expressions involving only geometric moments. No additional optimization over pose parameters is necessary. This procedure has been applied both to atlas construction as atlas-based segmentation. Secondly, we model the probabilistic appearance for each voxel and each label using parametric or nonparametric distributions and the prior term is determined by applying the vote-rule. The appearance and shape models are built with the variability of proposed linear registration. We adopt a graph cut - Bayesian framework for implementing the atlas-based segmentation. Finally, we illustrate the benefits of our approach on the liver segmentation from CT images.

References

1. Cootes, T., Taylor, C., Cooper, D., Graham, J., et al.: Active shape models-their training and application. Computer Vision and Image Understanding 61, 38–59 (1995)
2. Leventon, M., Grimson, W., Faugeras, O.: Statistical shape influence in geodesic active contours. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1. IEEE Computer Society, Los Alamitos (1999/2000)
3. Tsai, A., Yezzi, A., Wells, W., Tempany, C., Tucker, D., Fan, A., Grimson, W., Willsky, A.: A shape-based approach to the segmentation of medical imagery using level sets. IEEE Transactions on Medical Imaging 22, 137–154 (2003)

4. Cremers, D., Rousson, M.: Efficient kernel density estimation of shape and intensity priors for level set segmentation. In: Suri, J.S., Farag, A. (eds.) Parametric and Geometric Deformable Models: An application in Biomaterials and Medical Imagery. Springer, Heidelberg (2007)
5. Wang, Q., Seghers, D., D'Agostino, E., Maes, F., Vandermeulen, D., Suetens, P., Hammers, A.: Construction and validation of mean shape atlas templates for atlas-based brain image segmentation. In: Christensen, G.E., Sonka, M. (eds.) IPMI 2005. LNCS, vol. 3565, pp. 689–700. Springer, Heidelberg (2005)
6. Park, H., Bland, P., Meyer, C.: Construction of an abdominal probabilistic atlas and its application in segmentation. IEEE Transactions on Medical Imaging 22, 483–492 (2003)
7. Greig, D., Porteous, B., Seheult, A.: Exact maximum a posteriori estimation for binary images. Journal of the Royal Statistical Society. Series B (Methodological) 51, 271–279 (1989)
8. Ishikawa, H.: Exact optimization for Markov random fields with convex priors. IEEE Transactions on Pattern Analysis and Machine Intelligence 25, 1333–1336 (2003)
9. Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P.: Multimodality image registration by maximization of mutual information. IEEE Transactions on Medical Imaging 16, 187–198 (2002)
10. Heikkila, J.: Pattern matching with affine moment descriptors. Pattern Recognition 37, 1825–1834 (2004)
11. Pei, S., Lin, C.: Image normalization for pattern recognition. Image and Vision Computing 13, 711–723 (1995)
12. Saad, A., Hamarneh, G., Moller, T.: Exploration and Visualization of Segmentation Uncertainty Using Shape and Appearance Prior Information. IEEE Transactions on Visualization and Computer Graphics 16 (2010)
13. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1222–1239 (2001)
14. Klein, S., Staring, M., Murphy, K., Viergever, M., Pluim, J.: elastix: a toolbox for intensity-based medical image registration. IEEE Transactions on Medical Imaging 29 (2010)
15. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence 26, 1124–1137 (2004)
16. van Ginneken, B., Heimann, T., Styner, M.: 3D segmentation in the clinic: A grand challenge. In: Proceedings of MICCAI Workshop on 3D Segmentation in the Clinic: A Grand Challenge, pp. 7–15 (2007)

Error Bounds on the Reconstruction of Binary Images from Low Resolution Scans

Wagner Fortes[1,2] and K. Joost Batenburg[1,3]

[1] Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[2] Mathematical Institute, Leiden University, The Netherlands
[3] IBBT-Vision Lab, University of Antwerp, Belgium

Abstract. In this paper, we consider the problem of reconstructing a high-resolution binary image from several low-resolution scans. Each of the pixels in a low-resolution scan yields the value of the sum of the pixels in a rectangular region of the high-resolution image. For any given set of such pixel sums, we derive an upper bound on the difference between a certain binary image which can be computed efficiently, and *any* binary image that corresponds with the given measurements. We also derive a bound on the difference between any two binary images having these pixel sums. Both bounds are evaluated experimentally for different geometrical settings, based on simulated scan data for a range of images.

Keywords: Image reconstruction, Error bounds, Binary image, Rectangular scan.

1 Introduction

Black-and-white images, also called *binary images*, occur in a wide range of imaging applications. In many such applications, the images are actually acquired as grey level images by a scanning device. When scanning text, for example, binary characters are often scanned by a grey level scanner. When taking pictures of numberplates using a low resolution digital camera, the structure of the binary characters may even be unrecognizable in the resulting grey level images. Another example can be found in the single-pixel camera, which has recently been proposed within the framework of compressive sensing. Instead of recording individual fine-resolution pixels, such a camera records the total intensity over various areas of the object being photographed [8,11].

If several such grey level images are available, each representing a low resolution scan of some unknown ”original” binary image, one can attempt to reconstruct the binary image by combining the information from multiple scans [2,5,6]. In particular if the relative position of the different scans is well-known, this may lead to a high quality reconstruction. However, if the number of low resolution images available is relatively small in comparison with the resolution needed to properly represent the binary image, this reconstruction problem can be highly underdetermined. In such cases, many binary images can exist that correspond with the same scanned grey level data. At present, no useful bounds are available that can guarantee that the reconstructed image is actually close to the unknown original image.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 152–160, 2011.

In a recent paper [1], the authors presented bounds for binary image reconstruction in *tomography* (i.e. from projection data) that allow to bound the error between any two binary solutions, and therefore the error between the reconstructed binary image and the unknown original image. The proposed methodology is quite general and can potentially be extended to other imaging problems. As an intermediate step towards a general framework for bounding errors in binary image reconstruction, we apply the key concepts here to the problem of reconstructing binary images from low resolution scans.

2 Notation and Concepts

Let $A \subset \mathbb{Z}^2$ be a finite set, called the *reconstruction area*. We consider the problem of reconstructing a binary image defined on A, represented by a function $F : A \to \{0, 1\}$. A high resolution binary image defined on A will be reconstructed from several low resolution scans. The value of each pixel in such a scan corresponds with the summed intensity over all pixels in the corresponding region of the binary image. For simplicity, we assume here that the boundaries of the low resolution pixels coincide exactly with pixel boundaries in the high resolution binary image. We call a set $S \subset A$ a *window* of the reconstruction area. Let $\mathcal{S} = 2^A$, the set of all windows of A. We call a set $\mathbf{S} \subset \mathcal{S}$ of windows a *partition* of A if
(i) $S \cap T = \emptyset$ for all $S, T \in \mathbf{S}$ and
(ii) $\bigcup_{S \in \mathbf{S}} S = A$. We are also interested in the subsets of $\mathcal{S}$ which satisfy the property (i) but do not necessarily satisfy (ii). Such a subset will be called a *partial partition*. For $S \subset A$, define

$$P_F(S) = \sum_{(i,j)\in S} F(i, j). \tag{1}$$

We refer to the values $P_F(S)$ as *window-sums*. Note that our model for computing the window sums does not take certain properties of the imaging system, such as the detector point spread function, into account. However, the proposed methodology can easily be extended to include such effects, as long as they are linear. The *reconstruction problem* consists of finding an image F that has prescribed window-sum for a set $\mathbf{S}$ of windows. The existence and uniqueness of the solution of the general reconstruction problem is not guaranteed, in general.

To simplify the notation, the reconstruction problem can be formulated using linear algebra notation, which will be used in the forthcoming sections. Since there is an one-to-one mapping, say χ, from A to $\{1, \ldots, n\}$, the image F can be represented as a vector $\boldsymbol{x} = (x_j) \in \mathbb{R}^n$, where $n = \#A$ is the cardinality of A. We refer to the entries of $\boldsymbol{x}$ as *pixels*. A *binary image* on A corresponds with a vector $\bar{\boldsymbol{x}} \in \{0, 1\}^n$.

For a given set $\mathbf{S} \subset \mathcal{S}$ and an image $\boldsymbol{x} \in \mathbb{R}^n$, the combined set of window sums results in a vector $\boldsymbol{p} = (p_i) \in \mathbb{R}^m$, where m represents the number of window-sums taken. As the operator $P_F(S)$ is linear, the mapping from an image to its window sums can be represented by a matrix $\boldsymbol{W} = (w_{ij}) \in \mathbb{R}^{m\times n}$, which we call the *scan matrix*. The entry w_{ij} represents the weight of the contribution of x_j to the window-sum i.

Then, the general reconstruction problem can be stated as finding a solution of the system

$$\boldsymbol{W}\boldsymbol{x} = \boldsymbol{p}$$

for given window-sum data $\boldsymbol{p}$. In the binary image reconstruction problem, one seeks a binary solution of the system. For a given scan matrix $\boldsymbol{W}$ and a window-sum vector $\boldsymbol{p}$, let $\mathcal{T}_{\boldsymbol{W}}(\boldsymbol{p}) := \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{W}\boldsymbol{x} = \boldsymbol{p}\}$, the set of all real-valued solutions corresponding with the given data, and let $\bar{\mathcal{T}}_{\boldsymbol{W}}(\boldsymbol{p}) := \mathcal{T}_{\boldsymbol{W}}(\boldsymbol{p}) \cap \{0,1\}^n$, the set of *binary solutions* of the system.

As the scan matrix is typically not a square matrix, and also does not have full rank, it does not have an inverse. We recall that the *Moore-Penrose pseudo inverse* of an $m \times n$ matrix $\boldsymbol{A}$ is an $n \times m$ matrix $\boldsymbol{A}^\dagger$, which can be uniquely characterized by the two geometric conditions

$$\boldsymbol{A}^\dagger \boldsymbol{b} \perp \mathcal{N}(\boldsymbol{A}) \quad \text{and} \quad (I - \boldsymbol{A}\boldsymbol{A}^\dagger)\boldsymbol{b} \perp \mathcal{R}(\boldsymbol{A}) \quad \forall \boldsymbol{b} \in \mathbb{R}^m,$$

where $\mathcal{N}(\boldsymbol{A})$ is the nullspace of $\boldsymbol{A}$ and $\mathcal{R}(\boldsymbol{A})$ is the range of $\boldsymbol{A}$, [4, page 15].

Let $\boldsymbol{x}^* = \boldsymbol{W}^\dagger \boldsymbol{p}$. Then $\boldsymbol{x}^*$ also has the property (see Chapter 3 of [3]) that it is the minimal Euclidean norm solution of the system $\boldsymbol{W}\boldsymbol{x} = \boldsymbol{p}$, if it exists. We call $\boldsymbol{x}^*$ the *central reconstruction* of $\boldsymbol{p}$. The central reconstruction plays an important role in the bounds we derive for the binary reconstruction problem.

The description of the general reconstruction problem given above is quite broad and we will now specify the scan model by which we define the scan matrix $\boldsymbol{W}$ and the window-sum vector $\boldsymbol{p}$, in order to model the problem of reconstructing high resolution images from low resolution scans.

Put $A = \{(i, j) \in \mathbb{Z}^2 | \ 0 \leq i < l, \leq j < h\}$. Let $1 \leq p \leq l$, and $1 \leq q \leq h$. For $0 \leq i < l$, $0 \leq j < h$, define a rectangular set of pixels of size $p{\times}q$ by

$$S^{p,q}_{i,j} = \{(i + c, j + r) | \ 0 \leq c < p, 0 \leq r < q\}.$$

Each pixel in a low resolution scan corresponds to a *window* in our framework. It provides information about the summed intensity in a rectangular set of pixels of the scanned high resolution image. Adjacent low resolution pixels are connected and do not overlap. For $0 \leq a < p$ and $0 \leq b < q$, define

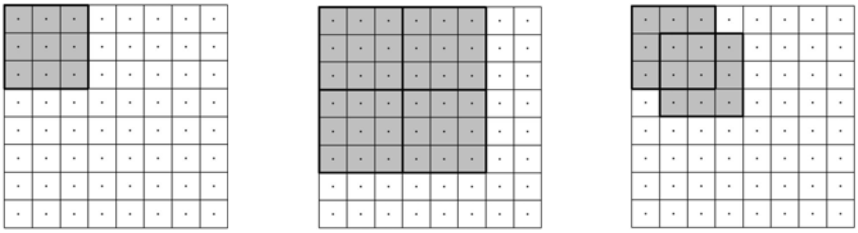
$$\mathbf{S}^{a,b} = \{S^{p,q}_{a+ip,b+jq} | \ a + ip < l, b + jq < h\}. \tag{2}$$

Each set $\mathbf{S}^{a,b}$ is a partial partition. Its elements correspond to pixels of the low resolution image. Let us now assume that several such low resolution images are available. Then the total set $\boldsymbol{S}$ of window-sums consists of the union of partial partitions $\mathbf{S}^d := \mathbf{S}^{a_d,b_d}$ for $d \in \{1, \ldots, k\}$. These concepts are illustrated in Fig. 1. Fig. 1a shows a single window $S_{a,b}$, whereas Fig. 1b shows the corresponding partial partition $\mathbf{S}^{a,b}$ formed by a tiling of its translates, where windows that cross the bounary of the image are not allowed. Fig. 1c shows two windows that are in separate partial partitions.

For $1 \leq d \leq k$, define the set of indices of the pixels x_j that were scanned by a partial partition $\mathbf{S}^d$ as $I_d := \{j | \ \chi^{-1}(j) \in \cup_{S \in \mathbf{S}^d} S\}$ and its complement $\bar{I}_d := \{1, \ldots, n\} \backslash I_d$.

As already mentioned, this linear scanning model can be modeled by a linear system of equations $\boldsymbol{W}\boldsymbol{x} = \boldsymbol{p}$. The matrix $\boldsymbol{W}$ and the window-sum $\boldsymbol{p}$ can be decomposed into k blocks as

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}^1 \\ \vdots \\ \boldsymbol{W}^k \end{pmatrix}, \quad \boldsymbol{p} = \begin{pmatrix} \boldsymbol{p}^1 \\ \vdots \\ \boldsymbol{p}^k \end{pmatrix}, \tag{3}$$



(a) Scan window

(b) A partial partition formed by translates of a scan window

(c) Scan windows for two different pairs (a, b)

Fig. 1. Rectangular scanning

where each block $\boldsymbol{W}^d$ $(d = 1, \ldots, k)$ represents the scanning of the image with a rectangular window as defined by $\mathbf{S}^d$ and each block $\boldsymbol{p}^d$ represents the corresponding window-sums $P_F(S)$ for $S \in \mathbf{S}^d$.

3 Error Bounds

Without loss of generality, we assume that all pixels in A are contained in at least one window. Clearly, no bounds can be given for those pixels that are not scanned at all, and they are removed from the analysis. As each set $\mathbf{S}^d$ samples a collection of disjoint subsets of A, the norm of the scanned binary image can be bound from above by the available window-sums:

Proposition 1. *Let $\bar{\boldsymbol{x}} \in \bar{\mathcal{T}}_{\boldsymbol{W}}(\boldsymbol{p})$. Then, $\|\bar{\boldsymbol{x}}\|_2^2 = \|\bar{\boldsymbol{x}}\|_1 \leq \|\boldsymbol{p}^d\|_1 + \#\bar{I}_d$ for all $1 \leq d \leq k$.*

The norm of any binary solution can therefore be estimated by summation of the window-sums in $\boldsymbol{p}^d$ and its accuracy increases with the number of scanned pixels included in the partial partition $\boldsymbol{S}^d$.

In the next Theorem we will use Prop. 1 to show that all binary solutions of the linear system $\boldsymbol{W}\boldsymbol{x} = \boldsymbol{p}$ have bounded distance to the central reconstruction $\boldsymbol{x}^*$.

Lemma 1. *Let $\bar{\boldsymbol{x}} \in \bar{\mathcal{T}}_{\boldsymbol{W}}(\boldsymbol{p})$ and $\boldsymbol{x}^* = \boldsymbol{W}^{\dagger}\boldsymbol{p}$. Put $R := \min_{1 \leq d \leq k} R_d$, where $R_d := \sqrt{\|\boldsymbol{p}^d\|_1 + \#\bar{I}_d - \|\boldsymbol{x}^*\|_2^2}$. Then, $\|\bar{\boldsymbol{x}} - \boldsymbol{x}^*\|_2 \leq R$.*

Proof. From the definition of $\boldsymbol{x}^*$ we have $(\bar{\boldsymbol{x}} - \boldsymbol{x}^*) \in \mathcal{N}(\boldsymbol{W})$, and $\boldsymbol{x}^* \perp (\bar{\boldsymbol{x}} - \boldsymbol{x}^*)$. Combining Pythagoras' theorem and Prop. 1 yields the theorem. □

We will now consider the image that is obtained by rounding each entry of $\boldsymbol{x}^*$ to the nearest binary value. Let $\langle \alpha \rangle = \min(|\alpha|, |\alpha - 1|)$ for $\alpha \in \mathbb{R}$, and put $U = \sqrt{\sum_{i=1}^{n} \langle x_i^* \rangle^2}$, i.e., the Euclidean distance from $\boldsymbol{x}^*$ to the nearest binary vector.

Let $\bar{\boldsymbol{r}} \in \{0, 1\}^n$ such that $\|\bar{\boldsymbol{r}} - \boldsymbol{x}^*\|_2 = U$, i.e., $\bar{\boldsymbol{r}}$ is among the binary vectors that are nearest to $\boldsymbol{x}^*$ in the Euclidean sense. If $R > U$ and $R - U$ is small, it is possible to say that a fraction of the rounded values are correct, i.e., to provide an upper bound on the *number* of pixel differences between any solution in $\bar{\mathcal{T}}_W(\boldsymbol{p})$ and $\bar{\boldsymbol{r}}$.

Suppose that $\bar{\boldsymbol{x}} \in \bar{\mathcal{T}}_W(\boldsymbol{p})$ and that $\bar{r}_i = 1$ whereas $\bar{x}_i = 0$. Note that we have $x^*_i \geq \frac{1}{2}$. Put $\tilde{\boldsymbol{r}} := \bar{\boldsymbol{r}}$ and then set $\tilde{r}_i$ to 0. We then have $||\tilde{\boldsymbol{r}} - \boldsymbol{x}^*||_2^2 = ||\bar{\boldsymbol{r}} - \boldsymbol{x}^*||_2^2 - |x^*_i - 1|^2 + |x^*_i|^2 = ||\bar{\boldsymbol{r}} - \boldsymbol{x}^*||_2^2 + 2x^*_i - 1$. Similarly, if $\bar{r}_i = 0$, then the squared Euclidean distance increases by $1 - 2x^*_i$ by setting pixel i to 1. Each time an entry i of $\bar{\boldsymbol{r}}$ is changed, the squared Euclidean distance to $\boldsymbol{x}^*$ increases by $b_i := |2x^*_i - 1|$.

As the Euclidean distance from $\boldsymbol{x}^*$ to $\bar{\boldsymbol{x}}$ is no greater than R, a bound can be derived on the maximal number of pixels in $\bar{\boldsymbol{r}}$ that must be changed to move from $\bar{\boldsymbol{r}}$ to $\bar{\boldsymbol{x}}$. Let us order the values b_i $(i = 1, \ldots, n)$ such that $b_i \leq b_{i+1}$ for $1 \leq i \leq n - 1$. Assuming that $\bar{\mathcal{T}}_W(\boldsymbol{p}) \neq \emptyset$, we have $R \geq ||\bar{\boldsymbol{r}} - \boldsymbol{x}^*||_2$ and the change of s entries of $\bar{\boldsymbol{r}}$ would increase the distance between $\bar{\boldsymbol{r}}$ and $\boldsymbol{x}^*$ such that $R^2 \geq ||\bar{\boldsymbol{r}} - \boldsymbol{x}^*||_2^2 + \sum_{j=1}^{s} b_j$.

Theorem 1. *Let $\bar{\boldsymbol{r}}$, $\bar{\boldsymbol{x}}$ and b_i $(i = 1, \ldots, n)$ be as defined above. Choose s such that*

$$\sum_{i=1}^{s} b_i \leq R^2 - ||\bar{\boldsymbol{r}} - \boldsymbol{x}^*||_2^2 < \sum_{j=1}^{s+1} b_j. \tag{4}$$

Then at most s pixels can have the wrong value in $\bar{\boldsymbol{r}}$ with respect to $\bar{\boldsymbol{x}}$ and at least $n - s$ pixels must have the correct value.

Proof. Due to the increasing order of the b_i's, changing more than s pixels in $\bar{\boldsymbol{r}}$ will result in a vector $\tilde{\boldsymbol{r}}$ for which $||\tilde{\boldsymbol{r}} - \boldsymbol{x}^*||_2 > R$, which cannot be an element of $\bar{\mathcal{T}}_W(\boldsymbol{p})$. □

Theorem 1 bounds the number of pixel differences between $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{r}}$, and between $\bar{\boldsymbol{y}}$ and $\bar{\boldsymbol{r}}$. When using these two bounds to determine an upper bound on the number of differences between $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$, we can assume that these two sets of pixel differences are disjoint, as otherwise the difference between $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ will only be smaller. This observation leads to the following corollary:

Corollary 1. *Let $\bar{\boldsymbol{r}}$ and b_i $(i = 1, \ldots, n)$ be as defined above. Let $\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}} \in \bar{\mathcal{T}}_W(\boldsymbol{p})$. Choose t such that*

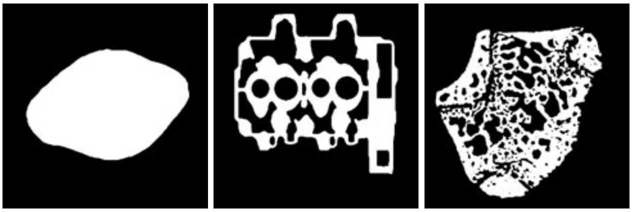
$$\sum_{i=1}^{t} b_i \leq 2(R^2 - ||\bar{\boldsymbol{r}} - \boldsymbol{x}^*||_2^2) < \sum_{j=1}^{t+1} b_j. \tag{5}$$

Then at most t pixels can be different between $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$.

4 Experiments and Results

A series of experiments was performed to investigate the of the bounds given in Theorem 1 and Corollary 1, for several test images. The experiments are all based on simulated data obtained by computing the low-resolution scans of a series of test images (called *phantoms*), shown in Fig. 2. All phantoms have a size of 512×512 pixels.

In each experiment, the central reconstruction $\boldsymbol{x}^*$ was first computed using the CGLS algorithm [9]. Depending on the experiment, this computation took from a few seconds, up to around 50s on a standard PC. The binary vector $\bar{\boldsymbol{r}}$ was computed by rounding $\boldsymbol{x}^*$ to

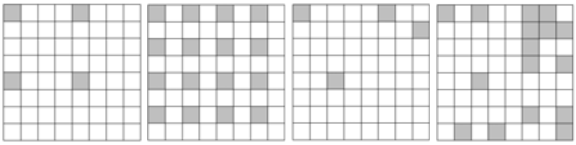


(a) Phantom 1 (b) Phantom 2 (c) Phantom 3

Fig. 2. Original phantom images used for the experiments

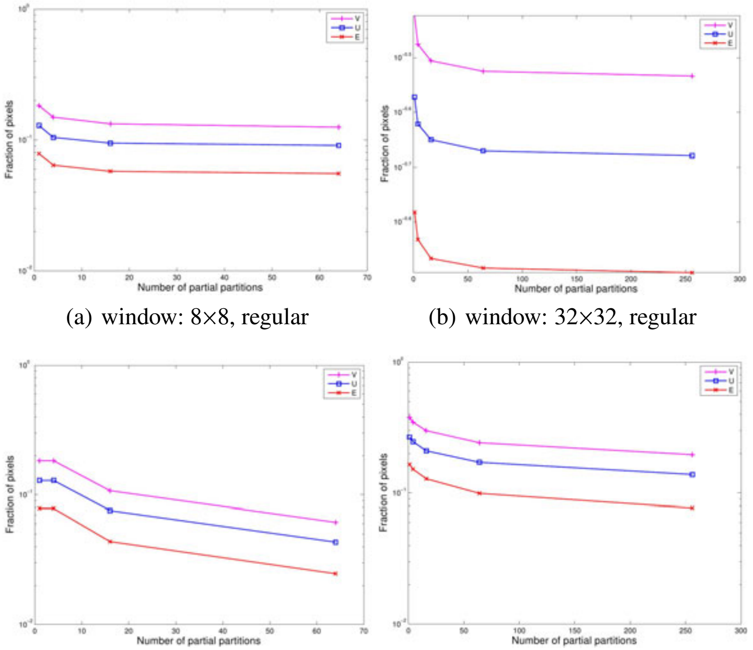
the nearest binary vector (choosing $\bar{r}_i = 0$ if $x_i^* = \frac{1}{2}$). The upper bound s from Theorem 1 on the number of differences between $\bar{\boldsymbol{r}}$ and the phantom image $\bar{\boldsymbol{x}}$ was then computed, followed by a bound on the fraction of pixel differences $U := \frac{s}{n}$, and the actual fraction of differences $E := \frac{e}{n}$, where e is the number of pixel differences between $\bar{\boldsymbol{r}}$ and $\bar{\boldsymbol{x}}$. The upper bound t from Corollary 1 on the number of differences any two binary solutions of $\boldsymbol{Wx} = \boldsymbol{p}$ was then computed, followed by the computation of the fraction of pixel differences $V := \frac{t}{n}$. Due to space limitations, we only show the results for Phantom 3. The results for the other two phantoms are in line with the observations made for the third phantom. In all experiments, a square window was used. Note that the position of a partial partition $\mathbf{S}^{a,b}$ with respect to the high resolution image is completely determined by the pair (a, b), which we call a *starting point*. Each low resolution image of the high resolution binary image corresponds to a different starting point. In the experiments, we distinguish between regularly and randomly distributed starting points, where the regular case corresponds to a low resolution scanner that is gradually shifted across the high resolution image, and the random case corresponds to a device that moves irregularly (or an object that moves in such a way); see Fig. 3. In Fig. 4, the three error measures V, U and E are plotted as a function of the number of starting points for window size of 8×8 and 32×32 and for both regularly and randomly distributed starting points. Note that for a larger window size, more starting points is required to obtain similar error bounds.

Various observations can be made from the graphs in Fig. 4. Even if the number of starting points is much smaller than the number of pixels in the window, meaning that the reconstruction problem is severely underdetermined, it is still possible to guarantee



(a) 4 points-regular (b) 16 points-regular (c) 4 points-random (d) 16 points-random

Fig. 3. Distribution of starting points in a first scan-window of size 8×8



(a) window: 8×8, regular

(b) window: 32×32, regular

(c) window: 8×8, random

(d) window: 32×32, random

Fig. 4. computed bounds as a function of the number of partial partitions for Phantom 3; V: bound on the distance between any two binary solutions from Cor. 1; U: bound on the distance between any binary solution and the rounded central reconstruction $\bar{\boldsymbol{r}}$ from Thm. 1; E: true error between the rounded central reconstruction and the binary phantom $\bar{\boldsymbol{x}}$

that only a limited fraction of pixels can be different between binary solutions. Although the given bounds U is clearly not sharp when compared to the real error E, rounding the central reconstruction yields a binary image that is in many cases guaranteed to be rather close to the original image. For example, for window size 8×8 and randomly distributed starting points, having just 16 low resolution images available (resulting in a system of equations that is underdetermined by a factor of 4) can still guarantee that the rounded central reconstruction is within 10% of the original binary image.

Fig. 5 illustrates the key concepts involved in the proposed bounds. The top row shows the central reconstruction for window sizes 8×8 and 32×32, with regularly and randomly distributed starting points. Here, the number of starting points is chosen as a fixed fraction of $\frac{1}{4}$ times the number of pixels in the window. In this way, all four reconstruction problems can be described by roughly the same number of equations. The middle row shows the difference images between the central reconstruction and the phantom, whereas the bottom row shows the difference images between the rounded central reconstruction and the phantom.

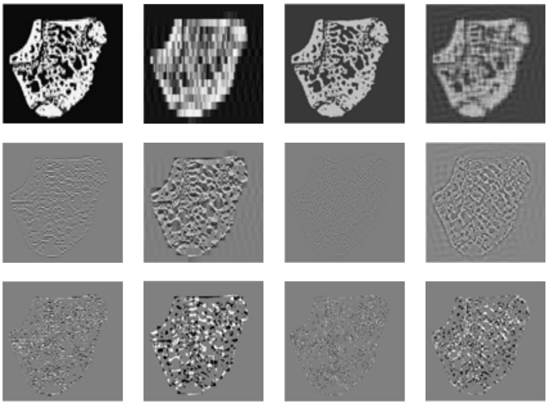


Fig. 5. Illustrations of the key concepts for Phantom 3; **From left to right**: 8×8 window, regular; 32×32 window, regular; 8×8 window, random; 32×32 window, random; **From top to bottom**: central reconstruction; difference between the central reconstruction and the phantom; difference between the rounded central reconstruction and the phantom

5 Outlook and Conclusions

In this article, we have presented general bounds on the accuracy of reconstructions of binary images from several low resolution graylevel scans, with respect to the unknown original image. The bounds can be computed efficiently and give guarantees on the number of pixels that can be different between any two binary reconstructions that satisfy given window-sums, and on the difference between a particular binary image, obtained by rounding the central reconstruction to the nearest binary vector, and any binary image satisfying the window-sums. The experimental results show that by using these bounds, one can prove that the number of differences between binary reconstructions must be small, even when the corresponding real-valued system of equations is severely underdetermined. This work represents an extension of the methodology set up in [1], which is a step towards a set of general bounds for binary image reconstruction problems that allow various forms of image sampling and incorporation of noisy measurements.

References

1. Batenburg, K. J., Fortes, W., Hajdu, L., Tijdeman, R.: Bounds on the difference between reconstructions in binary tomography. In: Proc. of the 16th IAPR International Conference on Discrete Geometry for Computer Imagery, pp. 369–380 (2011)
2. Batenburg, K.J., Sijbers, J.: Generic iterative subset algorithms for discrete tomography. Discrete Appl. Math. 157(3), 438–451 (2009)
3. Ben-Israel, A., Greville, T.N.E.: Generalized inverses: Theory and applications. Canadian Math. Soc. (2002)
4. Björck, Å.: Numerical methods for least square problems. SIAM, Linköping University, Sweden (1996)

5. Frosini, A., Nivat, M.: Binary matrices under the microscope: A tomographical problem. Theoretical Computer Science 370, 201–217 (2007)
6. Frosini, A., Nivat, M., Rinaldi, S.: Scanning integer matrices by means of two rectangular windows. Theoretical Computer Science 406, 90–96 (2008)
7. Hajdu, L., Tijdeman, R.: Algebraic aspects of discrete tomography. J. Reine Angew. Math. 534, 119–128 (2001)
8. Li, L., Stankovic, V., Stankovic, L., Li, L., Cheng, S., Uttamchandani, D.: Single pixel optical imaging using a scanning MEMS mirror. J. Micromech. Microeng. 21, 25022 (2011)
9. Saad, Y.: Iterative methods for sparse linear systems. SIAM, Philadelphia (2003)
10. Slavinsky, J., Laska, J., Davenport, M., Baraniuk, R.: The compressive multiplexer for multichannel compressive sensing. In: Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2011)
11. Wakin, M., Laska, J., Duarte, M., Baron, D., Sarvotham, S., Takhar, D., Kelly, K., Baraniuk, R.: An Architecture for Compressive Imaging. In: Proceedings of the International Conference on Image Processing (ICIP), pp. 1273–1276 (2006)

Tetrahedral Meshing of Volumetric Medical Images Respecting Image Edges

Michal Španěl, Přemysl Kršek, Miroslav Švub, and Vít Štancl

Department of Computer Graphics and Multimedia,
Faculty of Information Technology,
Brno University of Technology, Czech Republic
{spanel,krsek,svub,stancl}@fit.vutbr.cz

Abstract. In this paper, a variational tetrahedral meshing approach is used to adapt a tetrahedral mesh to the underlying CT volumetric data so that image edges are well approximated in the mesh. Afterwards, tetrahedra in the mesh are classified into regions whose image characteristics are similar. Three different clustering schemes are proposed to classify tetrahedra, while the clustering scheme viewing the mesh as an undirected graph achieved best results.

Keywords: CT data, image segmentation, surface reconstruction, 3D Delaunay triangulation, variational tetrahedral meshing, isotropic meshing, graph-cut segmentation.

1 Introduction

Medical imaging devices like the Computed Tomography (CT) produce volumetric image data detailing human anatomy. Such kind of data can be used for creation of 3D surface models of the anatomy, what is helpful for surgery planning and simulation.

This paper extends our previously described principles [1] where a tetrahedral mesh is used to partition the volumetric image data into regions (see Fig. 1). Process of the mesh construction respects detected image edges, hence surfaces of image regions are well approximated by the mesh and can be easily derived. Results further discussed in Sect. 3.1 show that reconstructed polygonal surfaces

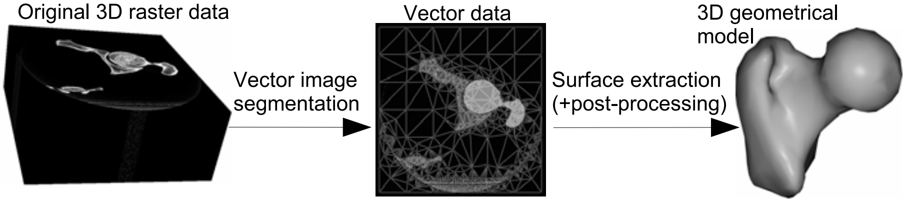

Original 3D raster data
Vector image segmentation
Vector data
Surface extraction (+post-processing)
3D geometrical model

Fig. 1. Surface reconstruction based upon the presented vector segmentation method

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 161–169, 2011.

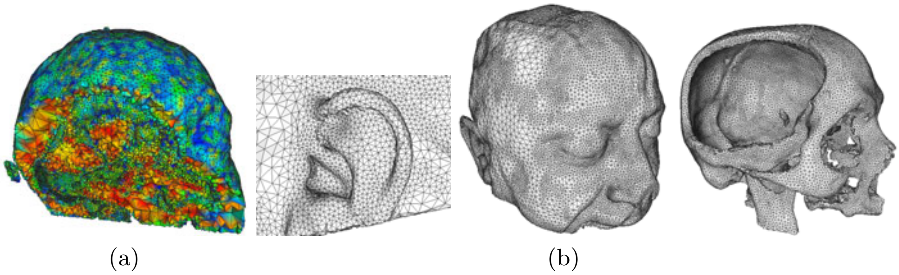


Fig. 2. Results of the proposed vector segmentation method: cut through the tetrahedral mesh (a); and surfaces extracted from the classified tetrahedral mesh (b)

are more accurate than those obtained by traditional techniques which typically process the surface meshes (e.g. smoothing and re-meshing) without any relationship to the original image data. In addition, meshes along with surfaces of all desired tissues are reconstructed at once.

A mesh generation [2] aims at tessellation of a bounded 3D domain Ω with tetrahedra. The *iso-surface stuffing* algorithm [3] was presented that fills an iso-surface with a uniformly sized tetrahedral mesh. The algorithm is fast and numerically robust. A variant of the algorithm creates a mesh with internal grading. However, the algorithm does not permit grading of both surface and interior tetrahedra.

Zhang *et al.* [4] presented an algorithm to extract adaptive and quality 3D meshes directly from volumetric image data. In order to extract tetrahedral (or hexahedral) meshes, their approach combines bilateral and anisotropic diffusion filtering of the original data, with octree subdivision and contour spectrum, iso-surface and interval volume selection.

Our meshing technique is partly based on the *variational tetrahedral meshing* (VTM) approach, proposed by Alliez *et al.* [5]. The VTM approach uses a simple quadratic energy to optimize vertex positions within the mesh and allows for global changes in mesh connectivity during energy minimization. This Delaunay-based meshing algorithm allows to create graded meshes (see Fig. 2), and defines a *sizing field* prescribing the desired tetrahedra sizes within the domain.

2 Delaunay-Based Vector Segmentation

The presented meshing technique is called *vector segmentation* (shortly *VSeg*) because it combines the Delaunay-based tetrahedral meshing as well as the image segmentation. It produces meshes whose tetrahedra are classified into particular regions – tissues. We have proposed the *VSeg* method as follows [1]:

- **Data preprocessing** – Noise reduction by means of 3D anisotropic filtering that performs piecewise smoothing of the image, while preserving edges.

- **3D edge and corner detection** – Points lying on region boundaries and strong edges are located.
- **Initial Delaunay triangulation** – Tetrahedral mesh is constructed from a sampled set of found edge points (Fig. 3) using the *Incremental method* [2].
- **Iterative adaptation** – The triangulation is adapted to the underlying image structure.
- **Mesh segmentation** – Final classification of tetrahedra into image regions.

Character and strength of edges differ between tissues. Thus, per concrete tissue, the image data are pre-processed using the *power-law contrast enhancement* technique to increase contrast of the desired tissue. Then, edges of the highlighted tissue are detected. In the end, all found edges from all different tissues are merged together. In our experiments, the well known *Canny edge detector* extended to 3D space has been used in each step.

Sizing Field. The sizing field enforces creation of larger tetrahedra inside image regions and smaller ones along region boundaries (i.e. image edges). We use the robust definition of presented by Alliez *et al.*. The sizing field is a function $h_P = \min_{S \in \delta\Omega}[Kd(S,P) + lfs(S)]$ defined at any point P of space that specifies the size of the elements in the mesh. Local feature size $lfs(P)$ at the domain boundary is defined as the distance $d(P, S_k(\Omega))$ to a *medial axis* of the domain. The parameter K controls gradation of the resulting field. In our case, instead of the conventional polygonal domain boundary, the sizing field respects found image edges:

1. Estimate 3D *distance transform* [6] from all image edges and find local maxima to identify the medial axis.
2. Evaluate local feature size $lfs(P)$ on image edges using the second distance transform propagating value from the medial axis.
3. Generate the sizing field distributing $lfs(P)$ from edges according to the function h_P.

2.1 Iterative Mesh Adaptation

During the adaptation, the following three steps are repeated. The idea is to grow the mesh (in the sense of vertices) until a predefined limit is reached:

- *Isotropic edge splitting* – creation of points along existing edges [2] introduces new points to the mesh,
- *Variational meshing* – optimization of the tessellation grid by means of the vertex moving,
- *Boundary refinement* – creation of new vertices along image edges to guarantee that all edges are well approximated by the tessellation grid.

Analogous to the original VTM approach [5], all boundary vertices are moved differently from the interior ones, however, the searching for boundary vertices is quite different. Each edge voxel V_i (the one where an image edge was detected) is

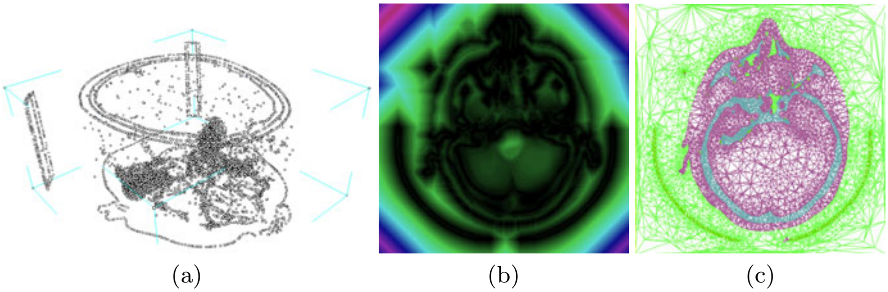


Fig. 3. Sampled set of points found by the edge and corner detection (a); orthogonal cut through the sizing field (b); and result of the tetrahedral mesh segmentation (c)

examined. Its nearest vertex S_j in the mesh is located, and the distance $d(V_i, S_j)$ as well as the coordinates of V_i (multiplied by the distance d) are accumulated at that vertex. Afterwards, vertices with a non–zero distance sum are moved to the average value they each have accumulated during the pass over all edge voxels – *Lloyd's algorithm* [7].

Boundary Refinement. We propose a novel mesh refinement technique to increase quality of the mesh in the sense of approximation of image edges. As other Delaunay refinement methods do, new vertices are added to the mesh.

1. For each edge voxel V_i:
   - Locate its nearest vertex S_j in the mesh and compare the distance $d(V_i, S_j)$ with the value stored in the corresponding accumulator. If the distance is smaller, change the value in the accumulator.
2. For all accumulators that contain a distance higher than d_{max}:
   - If the associated vertex is not itself located on an image edge, a new vertex is added to the mesh in place of the closest edge voxel.

Dealing with Slivers. Towards creation of a sliver-free mesh, the mesh is repeatedly tested for slivers, and new vertices lying in the center of sliver circumspheres are inserted to the mesh. If such addition does not eliminate the sliver, or generates new one, the vertex position is randomly perturbed. Such vertex perturbation continues until an optimal position is found.

2.2 Mesh Segmentation

Every tetrahedron t_i of the mesh is characterized by its feature vector that details the underlying image structure (mean value μ_{t_i} and variance σ_{t_i} of voxel intensity inside the tetrahedron, histogram of *Local Binary Patterns* [8], wavelet features, etc.). Concrete features must be chosen according to a specific task.

In our experiments, three different algorithms were used for the clustering of feature vectors into a certain number of classes:

- Fuzzy C-means (shortly *FCM*) algorithm [9],
- Gaussian Mixture Model optimized by the EM algorithm (*EM-GMM*) [10],
- Min-Cut/Max-Flow graph-based algorithm [11].

The feature extraction may be problematic if a tetrahedron is relatively small. Thus, we reject classification of small tetrahedra (limit of 15 voxels was chosen experimentally). These non-classified tetrahedra, that appear mostly near to region boundary, are assigned to particular regions in the next merging phase.

Agglomerative Merging. The agglomerative merging [12] sequentially reduces the number of regions (each region consists of one or more tetrahedra) by merging the best pair of adjacent regions among all possible pairs in terms of a given similarity measure. The similarity measure $S(r_j, r_i)$ is a function whose value is greater as the difference between two feature vectors t_i and t_j decreases:

$$S_\mu(r_j, r_i) = \frac{N_i + N_j}{N_i N_j} \exp(-\frac{1}{2\rho^2}|\mu_{r_i} - \mu_{r_j}|^2), \qquad S_\sigma(r_j, r_i) = \frac{\sigma_{r_i}\sigma_{r_j}}{\sigma^2_{r_{i,j}}}. \qquad (1)$$

where N_i is the volume of the region r_i in voxels, ρ affects sensitivity of the measure, and $\sigma_{r_{i,j}}$ is the variance of the intensity in a joint region $r_i \cup r_j$. Once the mesh is properly segmented, surface of any region r_k can be easily extracted. Boundary faces can be identified as faces between two different regions. The surface is closed and its mosaic conforms to the chosen parameters of the meshing.

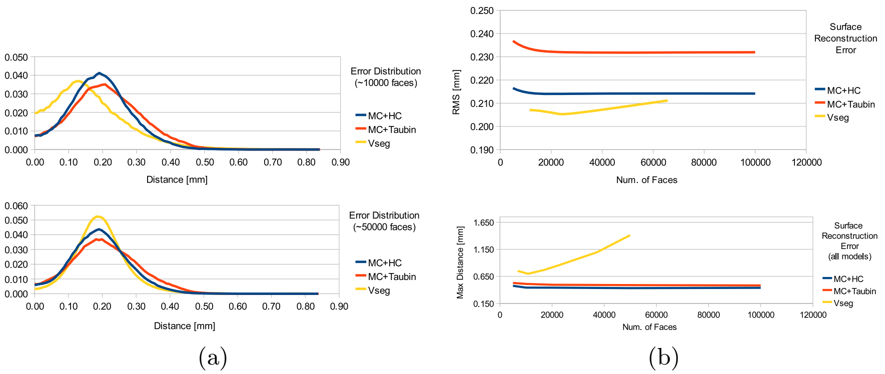
3 Experimental Results

The VSeg method was mainly designed for segmentation of volumetric medical images towards anatomical modeling of fundamental tissues (i.e. soft/bone tissues) and their surfaces.

3.1 Surface Accuracy

This evaluation compares surfaces produced by the vector segmentation against ones made by the traditional *Marching Cubes* (MC) method [13] followed by mesh smoothing and mesh decimation steps [14]. Since the smoothing is crucial for overall precision of the surface, two standard approaches were tested: Taubin's smoothing algorithm [15] that maintains volume of the mesh (*MC+Taubin*), and HC algorithm [16] that preserves sharp edges and corners in the mesh (*MC+HC*). After the smoothing, a variant of the *Quadric Edge Collapse* algorithm proposed by Garland and Heckbert [17] was used to reduce size of the mesh. Even thought these algorithms are not the best state of the art methods [18], they are well described in the literature and publicly available. Researchers may easily compare their results to this baseline.

An idea of the measurement was to rasterize basic solids into 3D raster, reconstruct surfaces from obtained artificial volumetric data, and evaluate an error between the reconstruction and the original surface. The approximation error



(a) (b)

Fig. 4. Histograms of the surface approximation error for two meshes with a different level of detail (a); and overall surface approximation error (b)

defined as the distance between corresponding sections of the meshes was estimated using the *Metro* [19] tool.

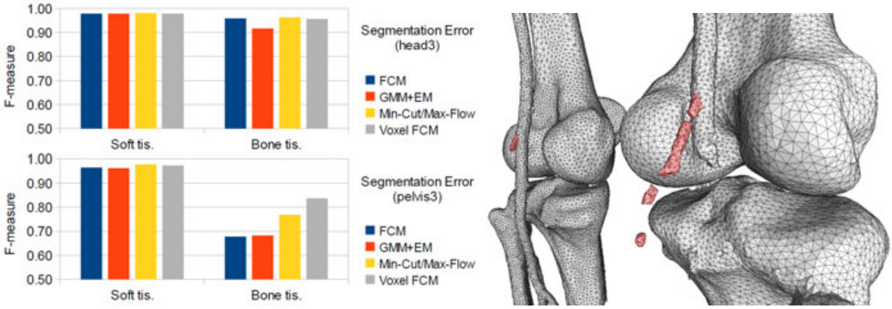
Fig. 4 shows histograms of error distribution for surface models of different level of detail. Apparently, the VSeg method outperforms both the smoothing based techniques. However, the difference is more evident for smaller meshes. The figure also illustrates the overall *RMS* (i.e. root mean square) error depending on the number of faces in the mesh. The same behavior can be seen. To obtain a more detailed surface, the minimal allowed edge length in the mesh L_{min} must be decreased. However, the resolution of the raster data is limited. Decreasing the L_{min} down to the real size of a single voxel causes that the relocation of vertices along image edges does not perform optimally.

The maximal error (maximal distance) between sampled points on compared surfaces, is greater for surfaces obtained by the VSeg method that generates meshes with almost regular tetrahedra. Close to the sharp surface edges, the final mesh approximates the surface very roughly because of the limitation of tetrahedra shape and the chosen minimal edge length.

3.2 Mesh Segmentation

The mesh segmentation was tested on several manually annotated CT data sets. Not unfrequently, the manual segmentation made by different people varies. The average error between two manual segmentations of the same data was about 0.96, measured by the *F-measure* of goodness (a perfect score is 1).

All parameters of the meshing phase were experimentally set to optimal values (most often $K = 1.5$ and $L_{min} = 1.5$). When compared to the manual segmentation, the VSeg method provides precise segmentation of the same quality as the voxel-based FCM clustering and the measured error is comparable to the variation of manual segmentation (Fig. 5). However, the error of the bone tissue segmentation significantly grows for the second dataset. Only the graph-based



(a) (b)

Fig. 5. Segmentation error of the VSeg method – three alternative clustering methods are compared to the straight *FCM* clustering of volumetric data (*voxel FCM*) (a); and surfaces reconstructed from pre-segmented data. In the red areas of the surface, small anatomical structures are weakly approximated because their size is relatively small compared to the prescribed minimal edge length.

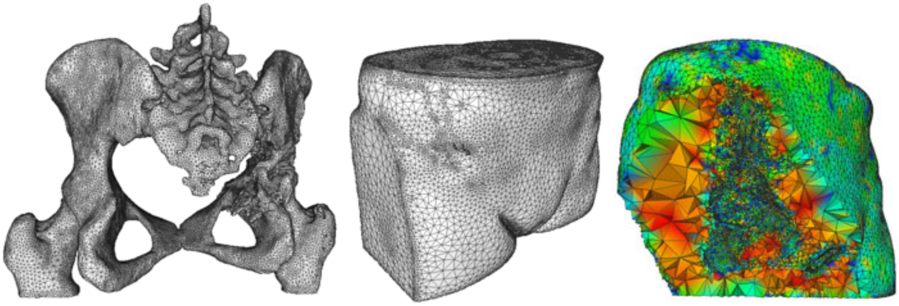


Fig. 6. Result of the VSeg method: CT data, $512x512x318$ voxels, resolution $0.63x0.63x0.70mm$; $K = 0.8$ and $L_{min} = 1.5mm$; soft tissues: 168308 faces; bone tissue: 238164 faces

Min-Cut/Max-Flow algorithm provides reasonable results because it takes spatial image structure into account. Due to the thickness of the cortical bone and resolution of the CT data, very thin edges are present in the image data which are practically undetectable by conventional edge detection techniques. Therefore, such edges are not well approximated during the meshing process which causes more errors.

All phases of the algorithm take approximately $25 - 50$ minutes on a standard PC with Intel 2.54GHz processor depending on a concrete size of the data and specific parameters of the meshing algorithm. The meshing itself consumes approximately $40 - 65\%$ of the total time. The MC–based techniques are able to reconstruct surfaces in a much less time – just about minutes. However, beside the surface, the VSeg method produces more comprehensive representation of the original image data – tetrahedral mesh, and it provides segmentation of the data at the same time.

4 Conclusions

The paper presents a technique for meshing of volumetric medical images aimed at surface reconstruction of fundamental human tissues (e.g. bone tissue). This *vector segmentation* technique is based on the 3D Delaunay triangulation. Such direct meshing of volumetric data appears to be more accurate approach than traditional techniques which start with the surface extraction followed by the decimation and smoothing without any relationship to the original image data.

Nevertheless, a more effective representation of the image structure is obtained. The mesh representation decreases complexity of the subsequent segmentation phase by means of processing a reduced number of tetrahedra instead of a large number of voxels. As an example, it was very difficult to segment the original voxel data using the Min-Cut/Max-Flow graph technique because of the large graph representation.

Results show that the vector segmentation can be successfully used for anatomical modeling of a human skull or soft tissues and the quality of reconstructed surfaces is sufficient. However, several inconveniences can be still found in the method. The original VTM approach produces well shaped tetrahedra inside the domain. However, slivers may appear close to the boundary. The same problem appears in case of the VSeg meshing method. Even thought the embedded sliver elimination algorithm removes a large number of poorly shaped tetrahedra, it does not ensure that all slivers will be successfully eliminated in a reasonable time. *J. Tournois* [20] presented a new modification of the VTM algorithm that particularly solves this problem and produces almost sliver free meshes.

Another disadvantage can be found in the edge detection step. In the future, we would like to incorporate constraints derived from edges and other image features directly into the original energy formulation of the VTM approach to obtain a more robust solution.

Acknowledgements. This work was partially supported by the BUT FIT grant FIT-S-10-2 and the research plan MSM0021630528. Many thanks to *O. Veksler* for his freely available implementation of the Min-Cut/Max-Flow algorithm [21] and thanks to *3Dim Laboratory s.r.o.* company for providing part of the test CT data.

References

1. Spanel, M., Krsek, P., Stancl, V.: Vector Segmentation of Volumetric Image Data: Tetrahedral Meshing Constrained by Image Edges. In: Proceedings of the 3rd International Joint Converence on Computer Vision, Imaging and Computer Graphics Theaory and Applications, pp. 134–138 (2010)
2. George, P.-L., Borouchaki, H.: Delaunay Triangulation and Meshing: Application to Finite Elements, 413 pages, (1998)

3. Labelle, F., Shewchuk, J.R.: Isosurface stuffing: fast tetrahedral meshes with good dihedral angles. ACM Trans. Graph. 26(3), 57 (2007)
4. Zhang, Y., Bajaj, C., Sohn, B.-S.: Adaptive and quality 3D meshing from imaging data. In: Proceedings of the Eighth ACM Symposium on Solid Modeling and Applications, pp. 286–291 (2003)
5. Alliez, P., Cohen-Steiner, D., Yvinec, M., Desbrun, M.: Variational tetrahedral meshing. ACM Trans. Graph. 24(3), 617–625 (2005)
6. Fabbri, R., Costa, L., da, F., Torelli, J.C., Bruno, O.M.: 2D Euclidean Distance Transform Algorithms: A Comparative Survey. ACM Computing Surveys 40(1), 1–44 (2008)
7. Du, Q., Emelianenko, M., Ju, L.: Convergence of the lloyd algorithm for computing centroidal voronoi tessellations. SIAM J. Numer. Anal. 44(1), 102–119 (2006)
8. Fehr, J., Burkhardt, H.: 3d rotation invariant local binary patterns. Pattern Recognition 29, 1–4 (2008)
9. Pham, D.L., Prince, J.L.: Adaptive fuzzy segmentation of magnetic resonance images. IEEE Transactions on Medical Imaging 18 (1999)
10. Ng, S.K., McLachlan, G.J.: On some variants of the em algorithm for fitting mixture models. Austrian Journal of Statistics 23, 143–161 (2003)
11. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(9), 1124–1137 (2004)
12. Kurita, T.: An efficient agglomerative clustering algorithm for region growing. In: Proc. of IAPR Workshop on Machine Vision Applications, pp. 210–213 (1991)
13. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. SIGGRAPH Comput. Graph. 21(4), 163–169 (1987)
14. Botsch, M., Pauly, M., Rossl, C., Bischoff, S., Kobbelt, L.: Geometric modeling based on triangle meshes. In: SIGGRAPH Course Notes (2006)
15. Taubin, G.: Geometric signal processing on polygonal meshes (2000)
16. Vollmer, J., Mencl, R., Mller, H.: Improved laplacian smoothing of noisy surface meshes. In: Computer Graphics Forum, vol. 18, pp. 131–138. Blackwell Publishing, Malden (1999)
17. Garland, M., Heckbert, P.S.: Surface simplification using quadric error metrics. In: SIGGRAPH 1997: Proceedings of the 24th annual conference on Computer graphics and interactive techniques, pp. 209–216 (1997)
18. Hildebrandt, K., Polthier, K.: Constraint-based fairing of surface meshes. In: Proc. of the Fifth Eurographics Symposium on Geometry Processing, pp. 203–212 (2007)
19. Cignoni, P., Rocchini, C., Scopigno, R.: Metro: measuring error on simplified surfaces. Computer Graphics Forum 17, 167–174 (1998)
20. Tournois, J., Srinivasan, R., Alliez, P.: Perturbing slivers in 3d delaunay meshes. In: Proceedings of the 18th International Meshing Roundtable, pp. 157–173 (2009)
21. Veksler, O.: Gcmex – matlab wrapper for graph cuts multi-label energy minimization (2010)

Measuring Shape Ellipticity

Mehmet Ali Aktaş and Joviša Žunić*

University of Exeter, Computer Science
Exeter EX4 4QF, U.K.
{M.A.Aktas,J.Zunic}@ex.ac.uk

Abstract. A new ellipticity measure is proposed in this paper. The acquired shape descriptor shows how much the shape considered differs from a perfect ellipse. It is invariant to scale, translation, rotation and it is robust to noise and distortions. The new ellipticity measure ranges over $(0, 1]$ and gives 1 if and only if the measured shape is an ellipse. The proposed measure is theoretically well founded, implying that the behaviour of the new measure can be well understand and predicted to some extent, what is always an advantage when select the set of descriptors for a certain application.

Several experiments are provided to illustrate the behaviour and performance of the new measure.

Keywords: Shape, shape descriptors, shape ellipticity, early vision.

1 Introduction

Image technologies have developed and used into many real-life applications: medical imaging [7], remote sensing [15], astronomy [1], etc. Different kind of objects, which appear on the images, should be classified, recognized or identified. An approach, to solve these problems, is to use pairwise features to match shapes from the same group/class. Such flexible matching technique could be inaccurate and computationally expensive. Another idea is to describe objects by using a set of numbers (a vector in R^d) and perform the searching in this space (e.g. in a subset of R^d). For such a description we need to extract object characteristics which can be reasonably easily and efficiently quantified by numbers. The shape is one of the object characteristics which enable a spectrum of numerical quantifications. Just to mention the colour and texture, as another object features which also enable quantifications with numbers. The shape allows a big diversity of numerical quantifications and, consequently, has a big differentiation capacity. Many shape descriptors were created and used. Some of them are quite generic: such as, Fourier descriptors [2] and moment invariants [5]. Alternatively, there are shape descriptors which use a single characteristic of shapes: Circularity [20], sigmoidality [11], linearity [13], rectilinearity [19], symmetry [16], etc. In this paper we deal with another global shape descriptor:

* J. Žunić is also with the Mathematical Institute of Serbian Academy of Science and Arts, Belgrade.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 170–177, 2011.

shape ellipticity. We define a new ellipticity measure which is well motivated, theoretically well founded, has a clear geometric meaning and has several desirable properties. The new measure is compared with several existing ellipticity measures, on small shape ranking, shape matching and shape classification tasks, in order to illustrate its behaviour and effectiveness.

The paper is organized as follows. Section 2 gives a brief review of the existing ellipticity measures used for the comparison with the new measure. The new measure is defined in Section 3. Experimental results are presented in Section 4. The conclusions are provided in Section 5.

2 Preliminaries

Ellipse is a basic shape widely applied to a vast range of image processing tasks involving not only man-made objects, but also natural forms. The problems like: How to determine the ellipse which fits best to the data considered, or how to evaluate how much a shape given differs from a perfect ellipse, have already been studied in literature [3,8,9,10,12]. Different techniques were employed – e.g. Discrete Fourier Transform [9], or affine moment invariants [10].

As expected, all the existing ellipticity measures have their own strengths and weaknesses, and it is not possible to establish a strict ranking among them. Measures which perform well in some tasks can have poor performance in others. In this paper we define a new ellipticity measure, and will compare its behaviour with several existing measures.

We begin a short overview, of the existing ellipticity measures, with a recent measure $\mathcal{E}_I(S)$ defined in [10]. The measure $\mathcal{E}_I(S)$ varies through the interval $[0, 1]$ and picks the value 1 when the considered shape S is an ellipse. The problem is that $\mathcal{E}_I(S) = 1$ does not guaranty (or at least this has not been proven) that the measured shape S is a perfect ellipse. Also, since $\mathcal{E}_I(S)$ is defined by using a projective invariant [4], it does not change the assigned ellipticity measure when an affine transformation is applied to the object considered. Of course, in some applications this property can be an advantage, but in some other applications it can be a disadvantage. $\mathcal{E}_I(S)$ uses the following affine moment invariant [4]:

$$I(S) = \frac{\mu_{20}(S) \cdot \mu_{02}(S) - \mu_{11}^2(S)}{\mu_{00}^4(S)} \tag{1}$$

and is defined as follows:

$$\mathcal{E}_I(S) = \begin{cases} 16 \cdot \pi^2 \cdot I(S) & \text{if} \quad I(S) \leq \dfrac{1}{16\pi^2} \\ \dfrac{1}{16 \cdot \pi^2 \cdot I(S)} & \text{otherwise.} \end{cases} \tag{2}$$

The quantities $\mu_{p,q}(S) = \iint_S \left(x - \frac{\iint_S x dx dy}{\iint_S dx dy} \right)^p \left(y - \frac{\iint_S y dx dy}{\iint_S dx dy} \right)^q dx dy$, appearing in (1), are well known as the centralized moments.

There are also some standard approaches which can be used to define an ellipticity measure. For example, the most common method [12] to determine an ellipse $E_f(S)$ which fits with a given shape S, also uses the moments for the computation. The axes of $E_f(S)$ are [12]:

$$major-axis: \quad \mu_{2,0}(S)+\mu_{0,2}(S)+\sqrt{4\cdot(\mu_{1,1}(S))^2+(\mu_{2,0}(S)-\mu_{0,2}(S))^2} \quad (3)$$

$$minor-axis: \quad \mu_{2,0}(S)+\mu_{0,2}(S)-\sqrt{4\cdot(\mu_{1,1}(S))^2+(\mu_{2,0}(S)-\mu_{0,2}(S))^2}. \quad (4)$$

The angle φ between the major axis of $E_f(S)$ and the x-axis is computed from

$$\tan(2\cdot\varphi)=\frac{2\mu_{11}(S)}{\mu_{20}(S)-\mu_{02}(S)}. \quad (5)$$

Now, we can define an ellipticity measure $\mathcal{E}_f(S)$ by comparing a given shape S and the ellipse $SE_f(S)$, which is actually the ellipse $E_f(S)$ scaled such that the area of S and the area of $E_f(S)$ coincide. A possible definition is:

$$\mathcal{E}_f(S)=\frac{Area(S\cap SE_f(S))}{Area(S\cup SE_f(S))}. \quad (6)$$

The angle φ, defined as in (5), is very often used to define the shape orientation [12]. The problem is that this method for the computation of the shape orientation fails in many situations, but also can be very unreliable [18]. Because of that, we modify the $\mathcal{E}_f(S)$ measure by replacing $SE_f(S)$ in (6) by rotating $SE_f(S)$ around the centroid for an angle θ which maximizes the area of $S\cap SE_f(S)$. If such a rotated ellipse $SE_f(S)$ is denoted by $SE_f(S(\theta))$ then we define a new ellipticity measure $\mathcal{E}_{fm}(S)$ as:

$$\mathcal{E}_{fm}(S)=\frac{Area(S\cap SE_f(S(\theta)))}{Area(S\cup SE_f(S(\theta)))}. \quad (7)$$

All three measures $\mathcal{E}_I(S)$, $\mathcal{E}_f(S)$, and $\mathcal{E}_{fm}(S)$, mentioned above, as well as the new ellipticity measure, which will be defined in the next section, are area based. This means that all the interior points are used for their computation. Because of that, we will say that all the shapes whose mutual set differences have the area equal to zero, are equal. For example, the shape of an open circular disc $\{(x,y) \mid x^2+y^2<1\}$ and the shape of the closed one $\{(x,y) \mid x^2+y^2\leq 1\}$ will be considered as equal shapes. Obviously, this is not a restriction in image processing tasks, but will simplify our proofs.

3 Main Result

In this section we give the main result of the paper. We define a new ellipticity measure and give some desirable properties of it. Throughout this section, it will be assumed, even not mentioned, that all appearing shapes have the unit area.

For our derivation we need an auxiliary ellipse $E(S)$ defined as

$$E(S) \;=\; \left\{(x,y) \mid \frac{x^2}{\rho(S)}+\rho(S)\cdot y^2\leq 1\right\}, \quad (8)$$

where $\rho(S)$ is the ratio between the major-axis and the minor-axis of S, defined as in (3) and (4) – $\rho(S)$ is also known as the shape elongation measure [12,14]. Notice that the area of $E(S)$ is 1. Now, we give the main result of the paper.

Theorem 1. *Let a given shape S whose area is 1 and whose centroid coincides with the origin. Let $S(\alpha)$ be the shape S rotated around the origin for an angle α, and let $Q(x,y) = \dfrac{x^2}{\rho(S)} + \rho(S)\cdot y^2$, for a shorter notation. Then:*

(a) $\iint_S Q(x,y)\ dx\ dy \ = \ \iint_{E(S)} Q(x,y)\ dx\ dy \quad \Rightarrow \quad S = E(S)$;

(b) $\min\limits_{\alpha\in(0,2\pi]} \iint_{S(\alpha)} Q(x,y)\ dx\ dy \ = \ \dfrac{1}{2} \quad \Leftrightarrow \quad S$ *is an ellipse.*

Proof. (a) Since all the points (x,y) satisfying $Q(x,y) = \dfrac{x^2}{\rho(S)} + \rho(S)\cdot y^2 \le 1$ are inside the ellipse $E(S)$ (see (8)) we deduce

$$(x,y)\in E(S) \quad \text{and} \quad (u,v)\notin E(S) \quad \Rightarrow \quad Q(x,y) < Q(u,v). \tag{9}$$

Now, by using the above implication, we derive

$$\iint_S Q(x,y)\ dx\ dy \ = \ \iint_{S\setminus E(S)} Q(x,y)\ dx\ dy \ + \ \iint_{S\cap E(S)} Q(x,y)\ dx\ dy \ \ge$$

$$\iint_{E(S)\setminus S} Q(x,y)\ dx\ dy \ + \ \iint_{E(S)\cap S} Q(x,y)\ dx\ dy = \iint_{E(S)} Q(x,y)\ dxdy. \tag{10}$$

Finally, the required implication (in *(a)*) follows from the fact that the equality $\iint_S Q(x,y)dxdy = \iint_{E(S)} Q(x,y)dxdy$ holds if and only if

$$\iint_{S\setminus E(S)} Q(x,y)\ dx\ dy \ = \ \iint_{E(S)\setminus S} Q(x,y)\ dx\ dy \ = \ 0$$

(a direct consequence of (9) and (10)) – i.e., if the shapes S and $E(S)$ are equal.

(b) This item follows from *(a)*, which actually says that $\iint_{S(\alpha)} Q(x,y)dxdy$ reaches the minimum possible value $1/2$ (notice $1/2 = \iint_{E(S)} Q(x,y)dxdy$ and see (10)) if there is an angle α such that $S(\alpha) = E(S)$. □

By the arguments of Theorem 1 we define the following ellipticity measure.

Definition 1. *Let a given shape S whose area is 1 and whose centroid coincides with the origin. The ellipticity $\mathcal{E}(S)$ of S is defined as*

$$\mathcal{E}(S) \ = \ \frac{1}{2}\cdot \frac{1}{\min\limits_{\alpha\in[0,2\pi]} \iint_{S(\alpha)} \left(\frac{x^2}{\rho(S)} + \rho(s)\cdot y(s)\right)\ dx\ dy}$$

where $\rho(S)$ is the elongation of S and $S(\alpha)$ denotes the shape S rotated around the origin for an angle α.

Now, we summarize desirable properties of $\mathcal{E}(S)$.

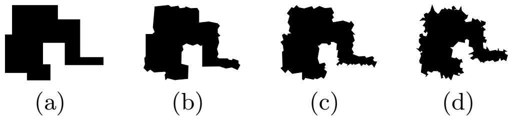
Theorem 2. *The ellipticity measure $\mathcal{E}(S)$ has the following properties:*

(a) $\mathcal{E}(S) \in (0,1]$;
(b) $\mathcal{E}(S) = 1$ *if and only if* S *is an ellipse;*
(c) $\mathcal{E}(S)$ *is invariant with respect translation, rotation and scaling transformations.*

Proof. The proof of *(a)* and *(b)* follows from Theorem 1. The proof of *(c)* follows directly from the definition. Basic calculus is sufficient for a formal proof. □

4 Experiments

In this section we perform several experiments to justify the effectiveness of the $\mathcal{E}(S)$ ellipticity measure and to compare it to the related measures $\mathcal{E}_f(S)$, $\mathcal{E}_{fm}(S)$, $\mathcal{E}_I(S)$). Notice that being area based, all these measures are robust (e.g. with respect to noise or to narrow intrusions) as it has been demonstrated in Fig.1. Even that the last shape has a big level noise added, the measured ellipticities do not change essentially.


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

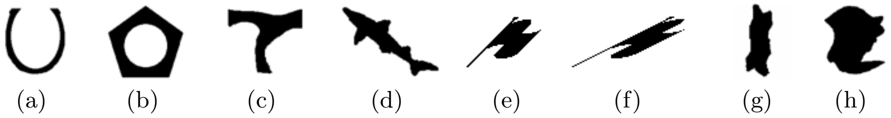
	(a)	(b)	(c)	(d)
$\mathcal{E}$	0.7484	0.7565	0.7617	0.7466
$\mathcal{E}_f$	0.6701	0.6786	0.6847	0.6668
$\mathcal{E}_{fm}$	0.6821	0.6969	0.7055	0.6929
$\mathcal{E}_I$	0.5622	0.5727	0.5813	0.5580

Fig. 1. Shapes with a different noise level added and their corresponded $\mathcal{E}$, $\mathcal{E}_f$, $\mathcal{E}_{fm}$, and $\mathcal{E}_I$ values

First Experiment. We start with examples in Fig.2. Eight random shapes are ranked in accordance with the increasing $\mathcal{E}(S)$ measure. The computed measures $\mathcal{E}(S)$, $\mathcal{E}_f(S)$, $\mathcal{E}_{fm}(S)$, and $\mathcal{E}_I(S)$ are in the table below the shapes. Notice that all measures can be understood as essentially different because they give different rankings. For example, if we consider the last 6 shapes the obtained rankings are:
$\mathcal{E} : (c)(d)(e)(f)(g)(h)$; $\mathcal{E}_f : (d)(c)(f)(e)(g)(h)$; $\mathcal{E}_{fm} : (c)(d)(f)(e)(g)(h)$; $\mathcal{E}_I : (c)(d)(e)(f)(h)(g)$ – i.e. all the rankings obtained are different.

The first shape in the same figure (Fig.2(a)) illustrates a big drawback of $\mathcal{E}_f(S)$ and $\mathcal{E}_{mf}(S)$. Both measures could assign the value 0 to the shapes with big holes or shapes whose centroid lies outside the shape. A consequence is that $\mathcal{E}_f(S)$ and $\mathcal{E}_{mf}(S)$ could not distinguish among such shapes. The new measure $\mathcal{E}(S)$ has no such a drawback and it does not take the value 0 for any shape.

The second shape in the same figure (Fig.2(b)) illustrates another drawback of the $\mathcal{E}_f(S)$ measure. I.e., it is well-known that $\mathcal{E}_f(S)$ cannot be applied to the N-fold rotationally symmetric shapes [18], because these shapes satisfy



Shape	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
$\mathcal{E}(S)$	0.1628	0.3011	0.6328	0.7039	0.7676	0.7691	0.9032	0.9033
$\mathcal{E}_f(S)$	0.0000	− − − −	0.5324	0.4838	0.6854	0.5326	0.7641	0.7752
$\mathcal{E}_{fm}(S)$	0.0000	0.0000	0.5374	0.6346	0.7426	0.7383	0.7612	0.7801
$\mathcal{E}_I(S)$	0.0266	0.0907	0.4010	0.5026	0.6120	0.6120	0.8305	0.8150

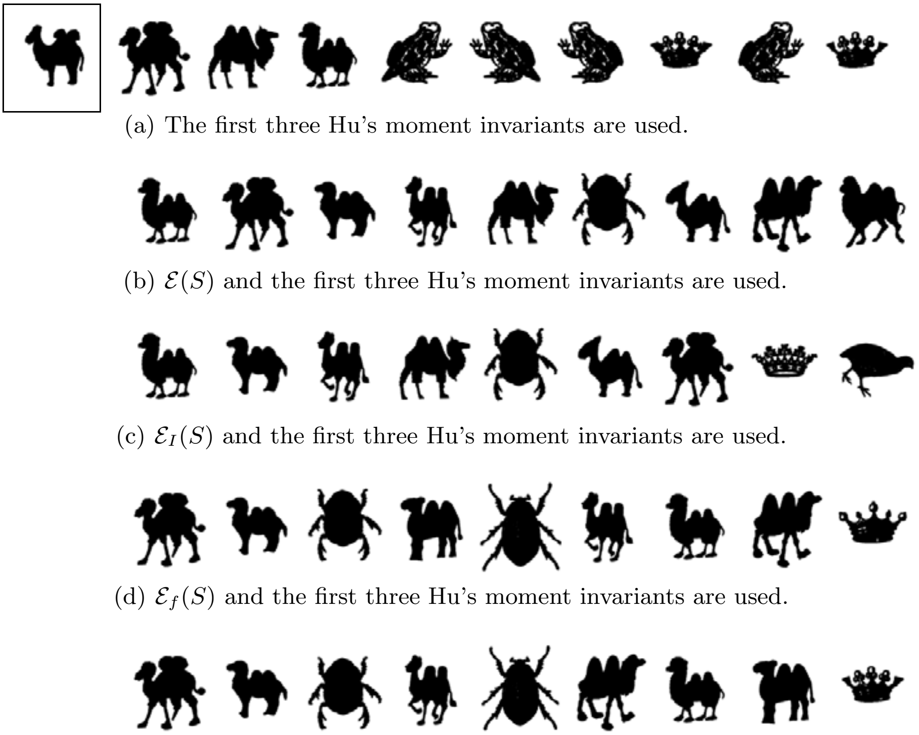
Fig. 2. Shapes are displayed in accordance with their increased $\mathcal{E}(S)$ measure

$\mu_{1,1}(S) = \mu_{2,0}(S) - \mu_{0,2}(S) = 0$ and, consequently, the orientation angle, defined as in (5), cannot be computed. The new measure $\mathcal{E}(S)$ does not have such a drawback. Notice that a big hole in the middle of the shape causes $\mathcal{E}_{fm}(S) = 0$ for this shape (as has already been discussed).

Finally, the shapes in Fig.2(e) and Fig.2(f) cannot be distinguished by the measure $\mathcal{E}_I(S)$ because it assigns the same value for all the shapes which are produced by affine transformations applied to a shape (as they are shapes in Fig.2(e) and Fig.2(f)). The new measure $\mathcal{E}(S)$ distinguishes among these shapes and this property can be an advantage in some applications. Of course, there are applications where this property is not preferred.

Second Experiment. In this experiment a shape matching task was performed. For this experiment the MPEG7 CE Shape-1 Part-B database was used. 200 images were chosen from 10 different classes (bat, camel, bone, crown, fork, frog, beetle, rat, horseshoe, bird) and the image "camel-7" was selected as the query image (the enclosed shape in Fig.3). In the first row the best 9 matches are displayed if the first three Hu moment invariants are used for the matching (3 of them were camels). In the next four task a single ellipticity measure form the set $\{\mathcal{E}(S),\ \mathcal{E}_f(S),\ \mathcal{E}_{fm}(S),\ \mathcal{E}_I(S)\}$ is used together with the first three Hu moment invariants and the best 9 matches are displayed in the corresponding rows. In all situation an improvement has been made, but the best improvement has been achieved once the new measure $\mathcal{E}(S)$ has been added to the set of the first three Hu moment invariants. In this case 8 out of 9 best matches were camels.

Third Experiment. The third task was to classify galaxies in two groups: spiral and elliptical. The data set consists of 104 images (100 × 100 pixels) and is originally used in [6]. Binary images were used for classification (i.e. images are thresholded before the classification, as shown in Fig.4). Four classification tasks were performed, each time by using a single ellipticity measure from the set $\{\mathcal{E}(S),\ \mathcal{E}_f(S),\ \mathcal{E}_{fm}(S),\ \mathcal{E}_I(S)\}$. The classification rates obtained are displayed in the table in Fig.4. It can be seen that the new ellipticity measure $\mathcal{E}(S)$ (75% classification rate achieved) has performed better than the measures $\mathcal{E}_f(S)$ (65.48%), $\mathcal{E}_{fm}(S)$ (67.31%), and $\mathcal{E}_I(S)$ (63.46%).



(a) The first three Hu's moment invariants are used.

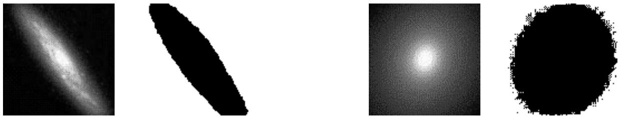
(b) $\mathcal{E}(S)$ and the first three Hu's moment invariants are used.

(c) $\mathcal{E}_I(S)$ and the first three Hu's moment invariants are used.

(d) $\mathcal{E}_f(S)$ and the first three Hu's moment invariants are used.

(e) $\mathcal{E}_{fm}(S)$ and the first three Hu's moment invariants are used.

Fig. 3. The enclosed query shape is in the first row. The best nine matches, for a different choice of shape descriptors used, are displayed in the corresponding rows.



(a) (b)

	Class. rate
$\mathcal{E}$	75.00%
$\mathcal{E}_f$	65.38%
$\mathcal{E}_{fm}$	67.31%
$\mathcal{E}_I$	63.46%

Fig. 4. Sample galaxy images with their shapes extracted by thresholding. The galaxy on the left (a) is spiral and the galaxy in (b) is elliptical.

5 Conclusion

A robust shape ellipticity measure is introduced in this paper. The new measure is invariant with respect to translation, rotation and scale transformations, ranges over $(0, 1]$ and gives 1 if and only if the measure shape is an ellipse. The new measure is theoretically well founded and has a clear geometric meaning - it indicates the difference between the considered shape and an ellipse. These two properties are very desirable because they give a prior indication about the measure suitability for the task planned to be performed. Notice that, for example, Hu's moment invariants (apart from the first one $\mu_{2,0}(S) + \mu_{0,2}(S)$, see [20]) do not have precise geometric interpretation – i.e., it is not clear which shapes maximize/minimize a certain invariant (for some related results see [17]).

Experiments provided illustrate theoretical observations and demonstrate applicability of the new ellipticity measure.

References

1. De Biase, G.A.: Trends in astronomical image processing. Computer Vision, Graphics, and Image Processing 43, 347–360 (1988)
2. Bowman, E.T., Soga, K., Drummond, W.: Particle shape characterization using Fourier descriptor analysis. Geotechnique 51, 545–554 (2001)
3. Fitzgibbon, A.M., Pilu, M., Fisher, R.B.: Direct least square fitting of ellipses. IEEE Transaction on Pattern Analysis and Machine Intelligence 21, 476–480 (1999)
4. Flusser, J., Suk, T.: Pattern recognition by affine moment invariants. Pattern Recognition 26, 167–174 (1993)
5. Hu, M.: Visual Pattern recognition by moment invariants. IRE Trans. Inf. Theory 8, 179–187 (1962)
6. Lekshmi, S., Revathy, K., Prabhakaran Nayar, S.R.: Galaxy classification using fractal signature. Astronomy and Astrophysics 405, 1163–1167 (2003)
7. Oliver, A., Freixenet, J., Martí, J., Pérez, E., Pont, J., Denton, E.R.E., Zwiggelaar, R.: A review of automatic mass detection and segmentation in mammographic images. Medical Image Analysis 14, 87–110 (2010)
8. Peura, M., Iivarinen, J.: Efficiency of simple shape descriptors. In: Arcelli, C., Cordella, L.P., Sanniti di Baja, G. (eds.) Aspects of Visual Form Processing, pp. 443–451. World Scientific, Singapore (1997)
9. Proffitt, D.: The measurement of circularity and ellipticity on a digital grid. Pattern Recognition 15, 383–387 (1982)
10. Rosin, P.L.: Measuring shape: ellipticity, rectangularity, and triangularity. Machine Vision and Applications 14, 172–184 (2003)
11. Rosin, P.L.: Measuring sigmoidality. Pattern Recognition 37, 1735–1744 (2004)
12. Sonka, M., Hlavac, V., Boyle, R.: Image Processing, Analysis, and Machine Vision. Thomson-Engineering (2007)
13. Stojmenović, M., Nayak, A., Žunić, J.: Measuring linearity of planar point sets. Pattern Recognition 41, 2503–2511 (2008)
14. Stojmenović, M., Žunić, J.: Measuring elongation from shape boundary. Journal Mathematical Imaging and Vision 30, 73–85 (2008)
15. Tran, A., Goutard, F., Chamaille, L., Baghdadi, N., Seen, D.: Remote sensing and avian influenza: A review of image processing methods for extracting key variables affecting avian influenza virus survival in water from Earth Observation satellites. Int. J. Applied Earth Observation and Geoinformation 12, 1–8 (2010)
16. Zabrodsky, H., Peleg, S., Avnir, D.: Symmetry as a continuous feature. IEEE Transactions on Pattern Analysis and Machine Intelligence 17, 1154–1166 (1995)
17. Xu, D., Li, H.: Geometric moment invariants. Pattern Recognition 41, 240–249 (2008)
18. Žunić, J., Kopanja, L., Fieldsend, J.E.: Notes on shape orientation where the standard method does not work. Pattern Recognition 39, 856–865 (2006)
19. Žunić, J., Rosin, P.L.: Rectilinearity measurements for polygons. IEEE Transactions on Pattern Analysis and Machine Intelligence 25, 1193–1200 (2003)
20. Žunić, J., Hirota, K., Rosin, P.L.: A Hu moment invariant as a shape circularity measure. Pattern Recognition 43, 47–57 (2010)

Robust Shape and Polarisation Estimation Using Blind Source Separation

Lichi Zhang and Edwin R. Hancock[⋆]

Department of Computer Science, the University of York,
{lichi.zhang,erh}@cs.york.ac.uk

Abstract. In this paper we show how to use blind source separation to estimate shape from polarised images. We propose a new method which does not require prior knowledge of the polariser angles. The two key ideas underpinning the approach are to use weighted Singular Value Decomposition(SVD) to estimate the polariser angles, and to use a mutual information criterion function to optimise the weights. We calculate the surface normal information using Fresnel equation, and iteratively update the values of weighting matrix and refractive index to a recover surface shape. We show that the proposed method is capable of calculating robust shape information compared with alternative approaches based on the same inputs. Moreover, the method can be applied when using uncalibrated polarisation filters. This is the case when the the subject is difficult to stabilse during image capture.

Keywords: Polarisation, SVD, Blind Source Separation, Shape Estimation.

1 Introduction

Three dimensional shape estimation from two dimensional brightness images is a key problem in computer vision, which has been approached using a number of approaches including shape-from-shading and photometric stereo. The use of polarisation although less widely studied, has also proved to be an effective method. Underpinning this approach is the Fresnel theory, which account for the way polarised light interacts with surfaces[13]. For dielectrics, polarisation may arise in two different ways. In the case of specular polarisation, initially polarised light is reflected in the specular direction. For diffuse polarisation, initially unpolarised light is refracted into the surface and the re-emitted light acquires a spontaneous polarisation. In both cases the zenith angle of the reflected or re-emitted light is constrained by the degree of polarisation, and the azimuth angle is constrained by the phase angle. Techniques derived from the Fresnel theory have been used for surface shape recovery[2][6][9].

In this paper we aim to use a Blind Source Separation (BSS) method to estimate polarisation state and recover shape from the result. Stated succinctly, we

[⋆] Edwin Hancock is supported by the EU FET project SIMBAD and by a Royal Society Wolfson Research Merit Award.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 178–185, 2011.

aim to extract the underlying source signal from a set of linear mixtures, where the mixing matrix is unknown[5]. The technique of BSS has found applications in the removal of reflections from transparent glass surfaces. Examples include the work of Fraid and Adelson [4] and Bronstein et al. [3].

Umeyama and Goldin [11] extend the method of Bronstein et al. using two polarised images and separate the diffuse and specular reflectance components. However, there are two main disadvantages which make it limited in applications. Firstly, it can only be used in the case where the brightness of specularities for the two input images are different, which is difficult to meet if the source light is unpolarised. Secondly, the information of phase angle is not considered, which is an important factor in the polarisation vision, therefore the separation result can not be accurately determined.

The principal contributions of this work are:

- We extend the work described in [11] and show how to accurately estimate polarisation state without prior knowledge of the input polariser orientations.
- Our method is based on a mutual information criterion, which is optimized with respect to determine the polariser angles and other parameters using Newton's method rather than exhaustive search, thus giving a relatively fast iterative procedure.
- Our method is also referred from the Fresnel theory. This leads to an iterative process that interleaves the processes of estimating shape based on the current polarisation state measurement, and updating the polarisation estimation based on the current shape estimate.
- We show how to use the proposed method to estimate refractive index, and prove that the results are physically reasonable.

2 Polarisation Estimation

In this section we show how to estimate polarisation state using blind source separation. We commence form a sequence of images collected with varying polariser angle. From these we aim to robustly estimate the three elements of the polarisation image, namely the mean intensity, degree of polarisation and phase. Wolff [12] gives a three images method for estimating the polarisation image using only three images collected with polariser orientations of 0, 45 and 90 degrees. The application is limited due to its specific requirement on polariser angles and poor robustness to noise.

A alternative approach available is to use the equation of Transmitted Radiance Sinusoid (TRS)[8]. Here more images can be used to estimate the polarisation image and eliminate the effects of noise. However, the method can not be applied when the subject is difficult to stabilse, which arouses problem of image alignment. Recently Saman and Hancock [10] introduced a method to estimate the polarisation image in a robust way, which improve the results obtained when the number of polarisation images is large.

When scattered light is transmitted through a linear polarizing filter, the intensity changes as the polariser angle θ is rotated. Let $I_{\max}$ denote the maximum brightness, $I_{\min}$ the minimum brightness and ϕ the phase angle (which corresponds to the angle of maximum transmission). The measured intensity follows the Transmitted Radiance Sinusoid (TRS) equation[8]:

$$\begin{aligned} I(\theta) &= \frac{(I_{\max}+I_{\min})}{2} + \frac{(I_{\max}-I_{\min})}{2}\cos(2\theta-2\phi) \qquad (1)\\ &= \frac{(I_{\max}+I_{\min})}{2} + \frac{(I_{\max}-I_{\min})}{2}\cos 2\theta \cos 2\phi + \frac{(I_{\max}-I_{\min})}{2}\sin 2\theta \sin 2\phi \ . \end{aligned}$$

Our method requires three $M \times N$ images captured under different polariser orientations θ_1, θ_2 and θ_3. Each image is converted into a long-vector of length MN. The long-vectors are the columns of the observation matrix $\boldsymbol{X}$. Consider the matrix

$$\begin{aligned} \boldsymbol{C} &= [(\frac{J_{max}+J_{min}}{2}), (\frac{J_{max}-J_{min}}{2}\cos 2\phi), (\frac{J_{max}-J_{min}}{2}\sin 2\phi)] \\ &= [C_a, C_b, C_c] \ . \end{aligned} \tag{2}$$

where J_{max} and J_{min} are long-vectors of length $M \times N$ that contain I_{max} and I_{min} as elements. We can determine the polarisation state via by solving the equation

$$\boldsymbol{X} = \boldsymbol{C}\boldsymbol{A}^T \ , \tag{3}$$

$$\boldsymbol{A} = \begin{bmatrix} 1 \cos 2\theta_1 \sin 2\theta_1 \\ 1 \cos 2\theta_2 \sin 2\theta_2 \\ 1 \cos 2\theta_3 \sin 2\theta_3 \end{bmatrix} \ . \tag{4}$$

To implement BBS, we commence by applying singular value decomposition (SVD) to the data matrix $\boldsymbol{X}$. We perform the operation without whitening as it will distort C_a [11]. The SVD of the data matrix $\boldsymbol{X}$ gives

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T \ . \tag{5}$$

where $\boldsymbol{U}$ is the $MN \times 3$ left eigenvector matrix, $\boldsymbol{D}$ the 3×3 diagonal matrix of singular values, and $\boldsymbol{V}$ the 3×3 right eigenvector matrix. To simplify the equation we let $\boldsymbol{P} = \boldsymbol{U}\ \boldsymbol{D}$. Referred from (3), it is noted that $\boldsymbol{P}$ is similar to $\boldsymbol{C}$ while as $\boldsymbol{V}$ is similar to $\boldsymbol{A}$. We define a 3×3 weighting matrix $\boldsymbol{W}$ satisfying $|\boldsymbol{W}| = 1$ such that

$$\boldsymbol{P} = \boldsymbol{C}\boldsymbol{W}^{-1} \ , \tag{6}$$

$$\boldsymbol{V} = \boldsymbol{A}\boldsymbol{W}^T \ . \tag{7}$$

From (7) we have $|\boldsymbol{W}| = |\boldsymbol{A}|/|\boldsymbol{V}| = 1$. As a result the elements of $\boldsymbol{P}$ can be determined by SVD, and the unknown elements of the matrix $\boldsymbol{A}$ by θ_1, θ_2 and θ_3. In fact, the value of θ_3 can be obtained when the values of θ_1 and θ_2 are known.

3 Shape Estimation

When unpolarised light is reflected from a surface, it will become partially polarised. As a result the measured brightness will be modulated by a linear polariser[13]. The variation of intensities range from I_{min} and I_{max}, and we can measure the degree of polarisation (DOP) denoted as ρ:

$$\rho = \frac{I_{max} - I_{min}}{I_{max} + I_{min}} \ . \tag{8}$$

According to the Fresnel theory[13] the polarisation is determined by the refractive index n and the zenith angle between the light source direction and surface normal (denoted by ψ)

$$\rho = \frac{(n - 1/n)^2 \sin^2 \psi}{2 + 2n^2 - (n + 1/n)^2 \sin^2 \psi + 4 \cos \psi \sqrt{n^2 - \sin^2 \psi}} \ . \tag{9}$$

When the refractive index n is known, the zenith angle ψ is determined by measurements of I_{max} and I_{min}. The azimuth angle of the surface normal can be decided using the estimated phase angle. However, a directional ambiguity must be resolved beforehand [7][1][6].

4 Iteration Process

Our aim is to estimate shape and polarisation information. There are three parameters to be estimated which are the polarised angles θ_1, θ_2 and the refractive n. Here we solve the problem by maximizing mutual information on estimated results and use the Newton Method for its rapid (quadratic) convergence.

Our mutual information criterion measures the similarity of the average intensity $\bar{I}^{(0)} = [I_{min}^{(0)} + I_{max}^{(0)}]/2$ and the diffuse reflectance component. This applies over most of the surface, except in the proximity of highlights. Ignoring specularities, we use Lambert's Law to link diffuse reflectance to the estimated surface normal direction. As a result we can write

$$I_{min}^{(0)} \propto \cos \psi^{(0)} \ . \tag{10}$$

We let $F^{(0)} = \cos \psi^{(0)}$, where the superscript is the iteration number. According to (10), for every pixel in the image the value of $F^{(0)}$ is monotone increasing with intensity value $\bar{I}^{(0)}$. As texture and highlight information is contained in $\bar{I}^{(0)}$, it is inadvisable to directly fit the values of $F^{(0)}$ to $\bar{I}^{(0)}$ using Lambert's Law. Instead, here we gauge their similarity using mutual information between their distributions. The aim is to find the set of parameters that maximise the distributional mutual information.

To compute the mutual information the probability density functions for the two measures together with their joint distribution function are required. We compute the distributions of $F^{(0)}$ and $\bar{I}^{(0)}$ using their associated normalised histograms, denoted as $x^{(0)}$ and $y^{(0)}$, respectively. Both histograms are quantised into L bins. The Shannon entropy for the probability density functions is:

$$H = -\sum_{i=1}^{L} p_i \log p_i \tag{11}$$

where p_i is the probability of density for bin i computed from the normalised histogram. For the two distributions the Shannon entropies are $H^{(0)}(x)$ and $H^{(0)}(y)$, respectively. To compute the the joint probability distribution, we construct the joint normalised histogram $z^{(0)}$. The Shannon entropy for the joint distribution is $H^{(0)}(x,y) = -\sum_{i=1}^{L}\sum_{j=1}^{L} z^{(0)}(i,j)\log z^{(0)}(i,j)$. Hence the mutual information between the distributions is given by

$$R(F^{(0)};\bar{I}^{(0)}) = H^{(0)}(x) + H^{(0)}(y) - H^{(0)}(x,y)\ . \tag{12}$$

Finally, the Newton method for updating the three parameters is written as

$$\Theta^{(t+1)} = \Theta^{(t)} - \gamma \boldsymbol{Q}[R^{(t)}]^{-1}\nabla R^{(t)}\ . \tag{13}$$

where $\Theta^{(m)} = (\theta_1^{(m)}, \theta_2^{(m)}, , n^{(m)})^T$, $\boldsymbol{Q}[R^{(t)}]$ is the Hessian of the error-function and ∇R its gradient. We initialize the parameters by setting $n^{(0)} = 1.4$ as this is typical of the materials studied, and set $\boldsymbol{A}^{(0)} = \boldsymbol{V}$.

5 Experiments

In this section we present the experiments with our new method for shape recovery and refractive index estimation, and compare it with alternatives. For each object studied, we collected three images using an unpolarised collimated source light placed in the direction of the camera (frontal illumination). A polariser is placed in front of the camera (a Nikon D200).

The first experiment conducted is to estimate shape and the polarisation state using the proposed method. Fig.1 shows the results. The three columns show the polarisation, phase angle, and the zenith angle for surface normal, respectively. The shape information in the third column is consistent with the subjective object shape, and any residual noise could easily be eliminated using a simple smoothing process [1]. This demonstrates that the estimation process works well without the information of concerning the polariser angles.

Next we explore the robustness of the method by randomly selecting three images from a longer sequence for the ball object, and check if the results remain unchanged. The polarisation orientations of the three sequences are: a)$30^o, 90^o, 150^o$, b)$60^o, 90^o, 120^o$, c)$0^o, 30^o, 120^o$. The results are presented in Fig.2. The polarisation and zenith angles are stable under the selection of different polariser angles, however there is an offset of 90 degrees in the phase angle. This suggests that we need to use constraints to consistently resolve the phase angle ambiguities.

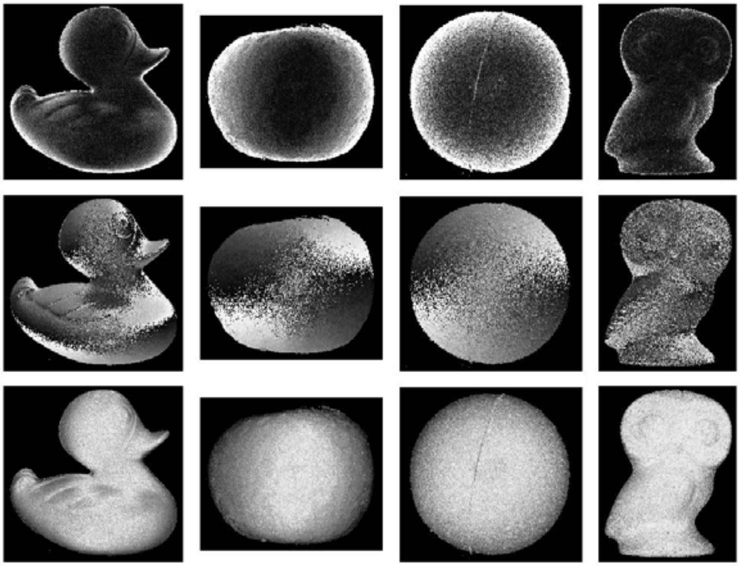


Fig. 1. The result for Shape from Polarisation, the first row is for DOP, the second for phase angle, and the third is for the result of zenith angle for the surface normal. The four columns are for different experiment objects which are the duck, the apple, the sponge ball and a plaster owl. The brightnesses for all graphs are adjusted to be displayed clearly.

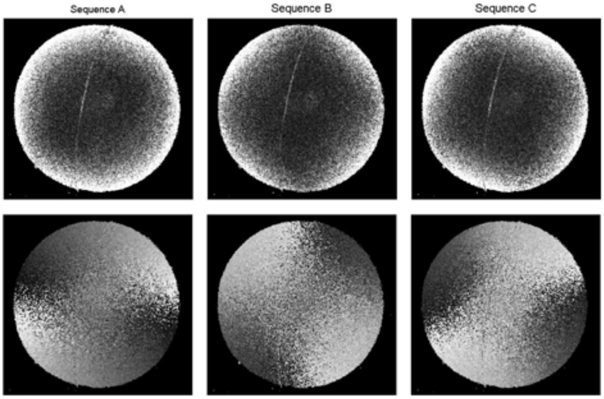


Fig. 2. The result for the estimation from different sequences, graphs for the first column stands for sequence a, the second column for sequence b and the third column for sequence c

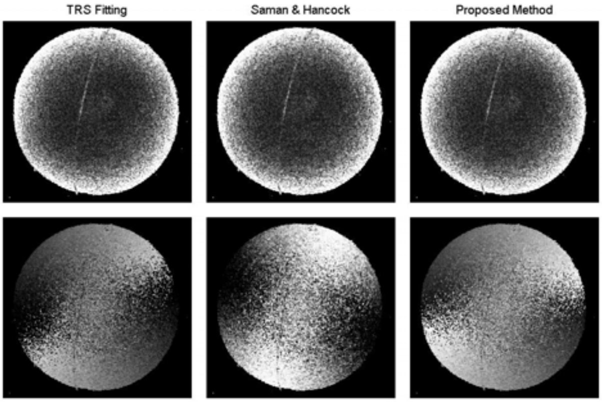


Fig. 3. The result for the three methods, the brightness have been adjusted

Table 1. The list of refractive index estimation results

Material	n_{ref}	n_{est}
Plastic	1.28	1.30
Apple	1.20	1.19
Plaster	1.46	1.22
Porcelain	1.51	1.53
Sponge	1.48	1.30

Then we compare our method with alternatives. We consider two methods, namely TRS fitting[8] and the method of Saman and Hancock [10]. Fig.3 shows the results for the three methods for the sponge ball, the first row is for DOP and the second the phase angle. There are no significant differences between the three methods for estimating DOP, while our proposed method performs best for the phase angle, which the distribution of the intensities is valid when representing the azimuth angles for the objects that ranges only from 0 to 180 degrees.

Finally, we explore the application of our method to refractive index estimation. In Table.1 we compare the measured refractive index values n_{est} with the tabulated values n_{ref} from five different materials. For the smooth surfaces, the results delivered by our method are all consistent with the tabulated results. However, for rough or indented surfaces such as plaster and sponge the results do not agree well. We attribute this to the effects of surface indentations which causes departures from Lambertian reflectance[1]. It is concluded that our method gives relatively accurate results for the refractive index of smooth objects.

6 Conclusion

In this paper we provide a new method for simultaneous shape from polarisation and refractive index estimation, which exploits blind source separation.

We demonstrate experimentally that it is both robust and reliable, and performs better than alternative methods. It can be used as the preprocessing step in shape segmentation, reflectance estimation and many other computer vision applications, especially when using non-calibrated polarisation filters. Future research will explore these applications.

References

1. Atkinson, G., Hancock, E.: Recovery of surface orientation from diffuse polarization. IEEE Transactions on Image Processing 15(6), 1653–1664 (2006)
2. Atkinson, G., Hancock, E.: Shape estimation using polarization and shading from two views. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(11), 2001–2017 (2007)
3. Bronstein, A., Bronstein, M., Zibulevsky, M., Zeevi, Y.: Sparse ICA for blind separation of transmitted and reflected images. International Journal of Imaging Systems and Technology 15(1), 84–91 (2005)
4. Farid, H., Adelson, E.: Separating reflections and lighting using independent components analysis. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 262–267 (1999)
5. Kisilev, P., Zibulevsky, M., Zeevi, Y., Pearlmutter, B.: Multiscal framework for blind source separation. Journal of Machine Learning Research 4, 1339–1363 (2004)
6. Miyazaki, D., Kagesawa, M., Ikeuchi, K.: Transparent surface modeling from a pair of polarization images. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(1), 73–82 (2004)
7. Miyazaki, D., Tan, R., Hara, K., Ikeuchi, K.: Polarization-based inverse rendering from a single view. In: International Conference on Computer Vision, vol. 2, pp. 982–987 (2003)
8. Nayar, S., Fang, X., Boult, T.: Separation of Reflection Components using Color and Polarization. International Journal of Computer Vision 21(3), 163–186 (1997)
9. Rahmann, S., Canterakis, N.: Reconstruction of specular surfaces using polarization imaging. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 149–156 (2001)
10. Saman, G., Hancock, E.: Robust Computation of the Polarisation Image. In: International Conference on Pattern Recognition, pp. 971–974 (2010)
11. Umeyama, S., Godin, G.: Separation of Diffuse and Specular Components of Surface Reflection by Use of Polarization and Statistical Analysis of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(5), 639–647 (2004)
12. Wolff, L.: Polarization vision: a new sensory approach to image understanding. Image and Vision computing 15(2), 81–93 (1997)
13. Wolff, L., Boult, T.: Constraining object features using a polarization reflectance model. IEEE Transactions on Pattern Analysis and Machine Intelligence 13(7), 635–657 (2002)

Hierarchical Representation of Discrete Data on Graphs

Moncef Hidane, Olivier Lézoray, and Abderrahim Elmoataz

Université de Caen Basse-Normandie, ENSICAEN, CNRS, GREYC Image Team
6 Boulevard Maréchal Juin, F-14050 Caen Cedex France
{moncef.hidane,olivier.lezoray,abderrahim.elmoataz-billah}@unicaen.fr

Abstract. We propose a new hierarchical representation of discrete data sets living on graphs. The approach takes advantage of recent works on graph regularization. The role of the merging criterion that is common to hierarchical representations is greatly reduced due to the regularization step. The regularization is performed recursively with a decreasing fidelity parameter. This yields a robust representation of data sets. We show experiments on digital images and image databases.

1 Introduction

Multilevel techniques are now well established in image processing. Generally, these techniques fall into two categories : multiresolution and multiscale. The former class yields a stack of successively blurred images and is well understood within the scale-space theory [1], while the latter is well formalized within the Multiresolution Analysis (MRA) framework [2]. MRA generally decomposes a signal into a coarse approximation and a detail or residual part. In graph theory, such an analysis is carried out through a decimation of the vertex set and a reconstruction. We propose in this paper a decimation procedure of functions whose support is the vertex set of a weighted graph. The weighted graph structure is of interest since its encapsulates pairwise interactions between discrete data instances. Furthermore, a graph structure can be obtained after sampling a continuous manifold. Our proposal is based on recent works on graph regularization and difference equations on graphs [3], [4]. The algorithm we propose produces a hierarchy of graphs and functions defined on their vertex sets. Starting with the initial data associated to a graph structure, successive coarsening procedures are applied. The coarsening is based on a preliminary graph partitioning which is mainly driven by a discontinuity-preserving graph regularization. The use of graph regularization yields a more robust representation. We show the applicability of our proposal to digital image hierarchical representation. In this case, the resulting representation can be seen as new adaptive irregular pyramidal representation of images [5]. The paper is organized as follows: in Section 2 we recall the graph regularization framework and present the algorithm we use. Section 3 details the different steps that lead to the representation we seek: regularization, grouping and coarsening. We present experiments and conclude in Section 4.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 186–193, 2011.

2 Graph Total Variation

2.1 Definitions

Throughout this section we assume that we are given a weighted graph $G = (V, E, w)$ consisting of a vertex set V, and an edge set $E \subseteq V \times V$. The nonnegative weight function $w : E \to \mathbb{R}^+$ is supposed symmetric: $w(\alpha, \beta) = w(\beta, \alpha)$ for all $(\alpha, \beta) \in E$. For a given graph edge $(\alpha, \beta) \in E$, the quantity $w(\alpha, \beta)$ represents a similarity or proximity measure between the vertices α and β. This measure is usually computed as a decreasing function of a prior distance measure. For $\alpha, \beta \in V$ we denote $\alpha \sim \beta$ if $(\alpha, \beta) \in E$. The graphs we consider in this paper are undirected with no self loops.

We denote by $\mathcal{H}(V)$ the set of functions that assign a real value to each vertex of the graph G and $\mathcal{H}(E)$ the set of functions that assign a real value to each edge. The sets $\mathcal{H}(V)$ and $\mathcal{H}(E)$ are equipped with the standard inner products denoted $\langle ., . \rangle_{\mathcal{H}(V)}$ and $\langle ., . \rangle_{\mathcal{H}(E)}$. The graph difference operator, $d_w : \mathcal{H}(V) \to \mathcal{H}(E)$, is defined as: $(d_w f)(\alpha, \beta) = \sqrt{w(\alpha, \beta)}\,(f(\beta) - f(\alpha)), \quad f \in \mathcal{H}(V), (\alpha, \beta) \in E$. The graph divergence operator, $\mathrm{div}_w : \mathcal{H}(E) \to \mathcal{H}(V)$, is related, as in the continuous setting, to the adjoint of d_w:

$$\langle d_w f, G \rangle_{\mathcal{H}(E)} = -\langle f, \mathrm{div}_w G \rangle_{\mathcal{H}(V)}, \quad \text{for all } f \in \mathcal{H}(V),\, G \in \mathcal{H}(E). \tag{1}$$

Its expression is given by: $(\mathrm{div}_w G)(\alpha) = \sum_{\beta \sim \alpha} \sqrt{w(\alpha, \beta)}(G(\alpha, \beta) - G(\beta, \alpha))$. For $p \in \mathcal{H}(E)$ and $\alpha \in V$, we denote $|p|_\alpha = \sqrt{\sum_{\beta \sim \alpha} p(\alpha, \beta)^2}$. A path joining two vertices α, $\beta \in V$ is a sequence of vertices $(\gamma_1, \ldots, \gamma_n)$ such that $\gamma_1 = \alpha$, $\gamma_n = \beta$ and $(\gamma_i, \gamma_{i+1}) \in E$, $i = 1, \ldots, n-1$.

2.2 Minimization

Let $f \in \mathcal{H}(V)$ associated with a graph structure G. The graph total variation (TV) of f is defined as: $TV_w(f) = \sum_{\alpha \in V} \sqrt{\sum_{\beta \sim \alpha} w(\alpha, \beta)(f(\alpha) - f(\beta))^2}$. Let $f_0 \in \mathcal{H}(V)$ be a possibly noisy data. In order to smooth f_0 we seek the minimum of the following functional: $E(f; f_0, \lambda) = TV_w(f) + \frac{\lambda}{2} \sum_{\alpha \in V} (f(\alpha) - f_0(\alpha))^2$. The parameter λ controls the amount of smoothing being applied to f_0.

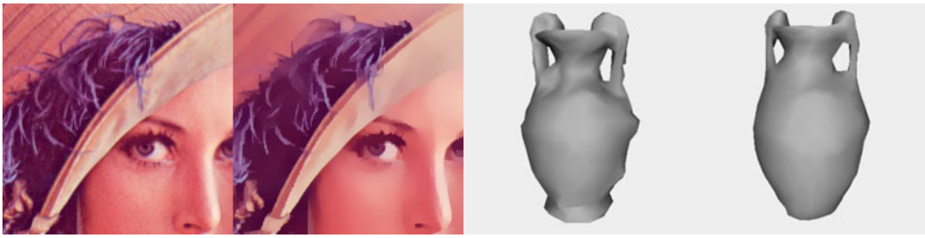
Functional E corresponds to the particular case of $p = 1$ in the family of functionals introduced in [3]. It has been studied in [6] and has found numerous applications in image and mesh processing [3], and data filtering [4]. In [7], the authors propose the adapt the penalization to the topology of the underlying function. The same approach has been taken in [8] for motion deblurring. Recently, [9] used functional E as a tool to generate an inverse scale space representation of discrete data on graphs.

Functional E is strictly convex but nonsmooth. Rather than smoothing the penalty term and differentiating, we use an adaption of Chambolle's projection algorithm [10] to graphs of arbitrary topologies. This adaption first appeared in [6]. In our notations, the solution of $\min \{E(f; f_0, \lambda),\ f \in \mathcal{H}(V)\}$ is given by :

$f = f_0 - \lambda^{-1}\mathrm{div}_w(p^\infty)$, where p^∞ is the limit of the following fixed point iterative scheme:

$$\begin{cases} p^0 = 0\,, \\ p^{n+1}(\alpha,\beta) = \dfrac{p^n(\alpha,\beta) + \tau\,(d_w(\mathrm{div}_w p^n - \lambda f))\,(\alpha,\beta)}{1 + \tau|d_w(\mathrm{div}_w p^n - \lambda f)|_\alpha}\,, \quad (\alpha,\beta) \in E\,. \end{cases} \tag{2}$$

If $0 < \tau \leq \frac{1}{\|\mathrm{div}_w\|^2}$, where $\|\mathrm{div}_w\|$ is the norm of the graph divergence operator, then (2) converges [6]. We use algorithm (2) in the next section as a tool to detect the possible groupings at different scales. Figure 1 shows the grouping effect yielded by a regularization of a color image and a triangular mesh. One should notice the preservation of discontinuities in the results.



(a) (b)

Fig. 1. Grouping effect of TV regularization. (a): left figure: original image; right: result of color components regularization with $\lambda = 0.01$. (b): left : original triangular mesh; right : result of spatial coordinates regularization with $\lambda = 0.01$.

3 Hierarchical Representation

3.1 TV as a Tool for Graph Partitioning

We propose to use TV regularization as a tool to detect the possible partitions in a given graph. The regularization yields a more regular data with respect to the graph TV prior while staying close to the original observations. The degree of closeness is inferred from the parameter λ . Once the TV regularization has been performed, a partitioning can be obtained by considering an equivalence relation on the vertex set. Let $G_i = (V_i, E_i, w_i)$ denote a given weighted graph, $f_i \in \mathcal{H}(V_i)$ and $\lambda > 0$. Let f_i^* be the result of TV regularization of f_i with parameter λ. We associate with each vertex $\alpha \in V_i$ a feature vector $F_i(\alpha)$ whose components are based on f_i^*. For instance, in image processing, the feature vector $F_i(\alpha)$ could consist of the values of $(f_i^*(\beta), \beta \in \mathcal{N}(\alpha))$ where $\mathcal{N}(\alpha)$ is an image patch centered at α. Define the metric d on V_i as the Euclidean distance between feature vectors: $d(\alpha,\beta) = \|F_i(\alpha) - F_i(\beta)\|_2$, $\alpha, \beta \in V_i$. Consider the following binary relation $\mathcal{R}^\epsilon$ on V_i: $\alpha\,\mathcal{R}^\epsilon\,\beta$ if $\alpha = \beta$ or if there exists a path $(\gamma_1, \ldots, \gamma_n)$ joining α and β such that: $d(\gamma_j, \gamma_{j+1}) < \epsilon$, for all $j = 1, \ldots, n-1$, $\epsilon > 0$. The relation $\mathcal{R}^\epsilon$ is reflexive, symmetric (G_i is undirected), and transitive. It is an equivalence relation. The quotient set $V_i/\mathcal{R}^\epsilon$ of V_i by $\mathcal{R}^\epsilon$ is a partition of V_i: $V_i/\mathcal{R}^\epsilon = \{[\alpha], \alpha \in V_i\}$ and $[\alpha] = \{\beta \in V_i \ : \ \alpha \mathcal{R}^\epsilon \beta\}$. The partition yielded by $\mathcal{R}^\epsilon$ can be seen as a region growing algorithm. The strength of the grouping is controlled by the parameter ϵ . It is important to understand that TV regularization will simplify the initial data and similar vertices will become closer. This will enable us to keep the parameter ϵ fixed in contrast to region merging techniques that do rely on variable thresholds.

3.2 Graph Coarsening

Let $P^\epsilon(V_i) = \{P_{i,1}, \ldots, P_{i,n_i}\}$ denote the partition of V_i obtained through the equivalence relation $\mathcal{R}^\epsilon$. In the sequel, we call the elements of $P^\epsilon(V_i)$ parts of V_i. We construct a coarse graph $G_{i+1} = (V_{i+1}, E_{i+1}, w_{i+1})$ by aggregating the nodes belonging to each part. Let $\gamma \in V_{i+1}$ be a vertex in the coarse graph. We denote R^i_γ the set of vertices in V_i which have been aggregated into γ in V_{i+1}. Two nodes $\gamma_1, \gamma_2 \in V_{i+1}$ are connected by a coarse edge if there exits $\alpha_1 \in R^i_{\gamma_1}$, $\alpha_2 \in R^i_{\gamma_2}$ such that α_1 and α_2 are connected by a (fine) edge. In the latter case, we adopt the notation $R^i_{\gamma_1} \sim R^i_{\gamma_2}$.

In order to take account of the volumes of the parts obtained by the partitioning, the edges of the coarse graph should be weighted. We use the ratio-cut measure between two parts in the fine graph as the weight between their aggregates in the coarse graph: $w_{i+1}(\gamma_1, \gamma_2) = \frac{cut(R^i_{\gamma_1}, R^i_{\gamma_2})}{|R^i_{\gamma_1}|} + \frac{cut(R^i_{\gamma_1}, R^i_{\gamma_2})}{|R^i_{\gamma_2}|}$, where $cut(A, B) = \sum_{(a,b)\in A\times B} w(a, b)$ is the edge cut between A and B. Once the coarse graph $G_{i+1} = (V_{i+1}, E_{i+1}, w_{i+1})$ has been constructed, we define a new function $f_{i+1} \in \mathcal{H}(V_{i+1})$ by averaging the values of each part: $f_{i+1}(\gamma) = \frac{1}{|R^i_\gamma|} \sum_{\alpha \in R^i_\gamma} f_{i+1}(\alpha)$, $\gamma \in V_{i+1}$.

3.3 Recursive Construction of the Hierarchy

We have showed in the two previous sections how to construct a weighted coarse graph from an input graph and function pair. We now move on to see how this process can be repeated to generate a hierarchy of graphs.

A hierarchical representation can be obtained by varying the ϵ parameter. However, we do not follow this direction in this section and ϵ will remain fixed within all the hierarchy: the partitioning is induced by the TV regularization. We seek to adapt the different levels of the representation to the local properties of the data.

The hierarchical representation is based on recursive partitioning and coarsening as described above. In order to adapt to the local properties of data, the fidelity parameter λ should evolve through the hierarchy. In our case, λ should decrease through the coarsening process, favoring more regularity and less fidelity as the hierarchy evolves. In our experiments, we have chosen a dyadic progression $\lambda_{i+1} = \frac{\lambda_i}{2}$. The initial regularization is responsible for denoising the initial data. It yields a choice for the first fidelity parameter λ_0 which is set to $\frac{1}{\sigma^2}$ where σ^2 is the variance of the noise, which we suppose Gaussian (see [11]). The standard deviation can be estimated through the standard estimator: $\widehat{\sigma} = 1.4826\,\mathrm{MAD}(f_0(\alpha),\, \alpha \in V)$, where MAD is the median absolute deviation estimator [12]. Finally we summarize the algorithm:

Algorithm 1. Hierarchical representation of discrete data on graphs

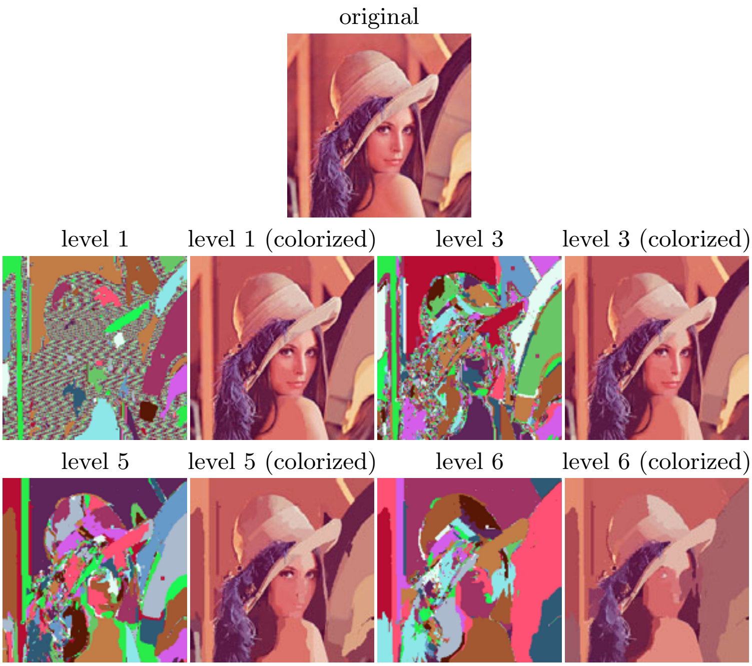
```
INPUT: $G_0 = (V_0, E_0, w_0)$, $f_0 \in \mathcal{H}(V_0)$, $\lambda_0 = \frac{1}{\sigma^2}$, $n \geq 1$, $\epsilon$ fixed
for $i = 0$ n do
    Regularization : $f_i^* \leftarrow$ arg min $\{E(f; f_i, \lambda_i),\, f \in \mathcal{H}(V_i)\}$
    Partitioning of $V_i$ based on $f_i^*$ through equivalence relation $\mathcal{R}^\epsilon$ to yield $P(V_i) = \{P_{i,1}, \ldots, P_{i,n_i}\}$
    Coarse nodes: Aggregate each part to yield $V_{i+1} = \{j_1, \ldots, j_{n_i}\}$
    Coarse edges : $E_{i+1} = \{e_{\alpha,\beta} \ : \alpha, \beta \in V_{i+1} \quad \text{and} \quad R^i_\alpha \sim R^i_\beta\}$
    Coarse weights: $w_{i+1}(\alpha, \beta) = \frac{cut(R^i_\alpha, R^i_\beta)}{|R^i_\alpha|} + \frac{cut(R^i_\alpha, R^i_\beta)}{|R^i_\beta|}$, $(\alpha, \beta) \in E_{i+1}$
    Coarse function $f_{i+1}(\alpha) = \frac{1}{|R^i_\alpha|} \sum_{\beta \in R^i_\alpha} f_i(\beta), \quad \alpha \in V_{i+1}$
    Update the fidelity parameter : $\lambda_{i+1} = \lambda_i / 2$
end for
```

4 Experiments and Conclusion

We begin by applying our approach to digital images. The algorithm we propose leads to a hierarchy of partitions of an image. Each pixel is represented by a vertex. In the experiments, we have chosen an eight connectivity graph. The edge weights are computed as follows: $w(\alpha, \beta) = \exp(-\frac{d^2(\alpha,\beta)}{\sigma^2})$, where d is the Euclidean distance between two RGB color vectors. The function to regularize is the one that assigns to each pixel its RGB color values. The regularization of multivalued functions is carried on each component but with a common total variation prior. The merging at the first stage is based on the distance between RGB color patches (5x5 in our case). At the following stages, its is based on vertex-wise distance. In all cases, the parameter ϵ was set to one. Figures 2 and 3 show the result of the regions obtained as well as their colorizations based on the original image.

We also show an application of our approach to image databases. Here each vertex represents a given image. The edges are obtained by considering a nearest neighbor graph (NNG) weighted with $w = 1/d$. The number of neighbors was set to 7. Figure 4 shows the hierarchy obtained. One should notice that the graph structure evolves as well as the image data. This yields simplification as well as decimation.

As Figures 2 and 3 show, its is difficult to get rid of outlier pixels in the first levels. Its seems interesting to adopt an approach based on concentration inequalities as used in [13] to replace the equivalence relation $\mathcal{R}^\epsilon$. This will be the subject of a future work.

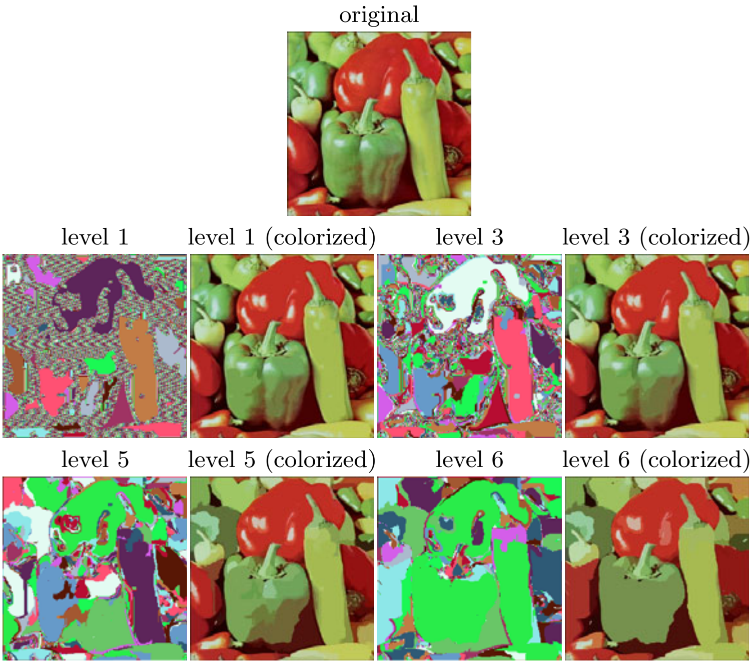


original

level 1 level 1 (colorized) level 3 level 3 (colorized)

level 5 level 5 (colorized) level 6 level 6 (colorized)

Fig. 2. Hierarchy of partitions and corresponding colorizations. Levels 1, 3, 5 and 6.



original

level 1 level 1 (colorized) level 3 level 3 (colorized)

level 5 level 5 (colorized) level 6 level 6 (colorized)

Fig. 3. Hierarchy of partitions and corresponding colorizations. Levels 1, 3, 5 and 6.

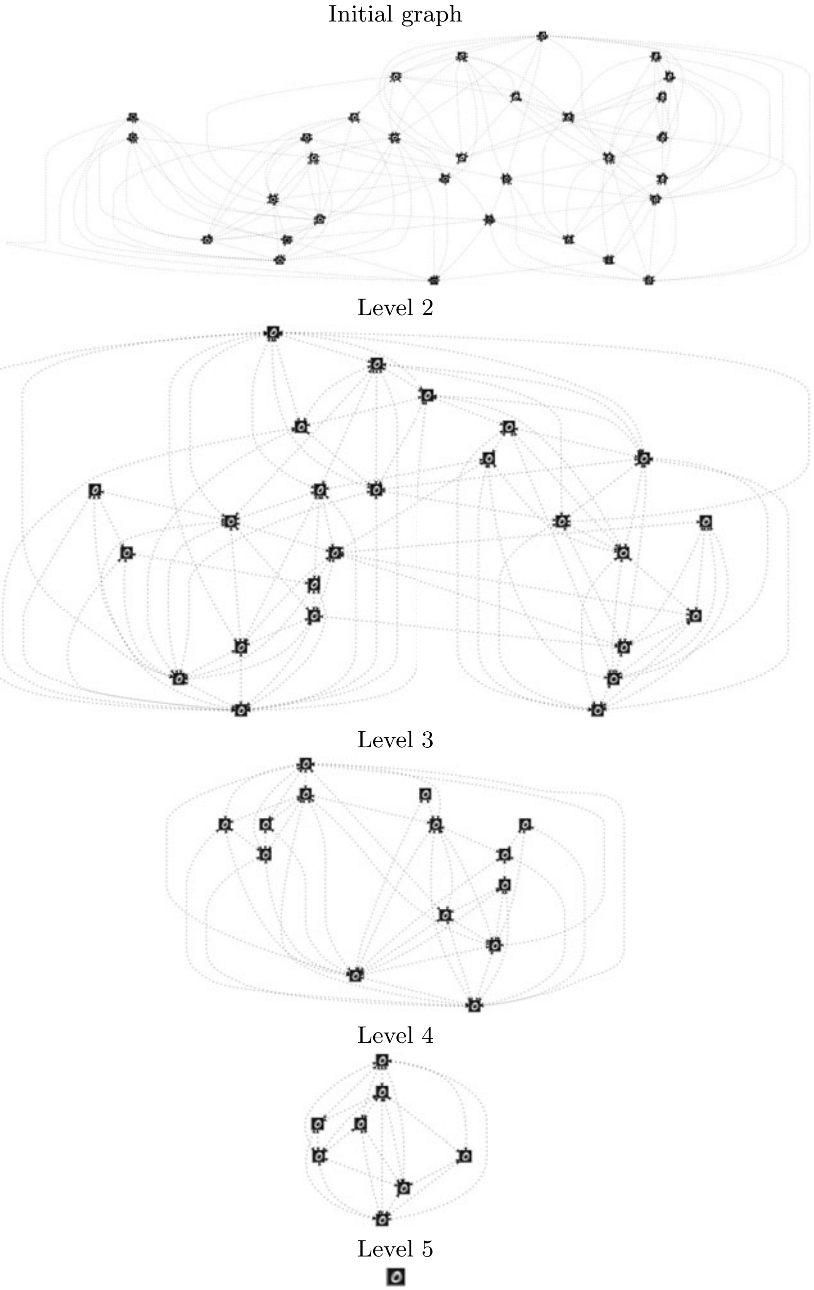


Fig. 4. 0-digits database hierarchical representation

References

1. Weickert, J.: Anisotropic Diffusion in Image Processing. ECMI Series. Teubner (1998)
2. Mallat, S.: Wavelet Tour of Signal Processing, 3rd edn. The Sparse Way. Academic Press, London (2008)
3. Elmoataz, A., Lezoray, O., Bougleux, S.: Nonlocal discrete regularization on weighted graphs: a framework for image and manifold processing. IEEE Transactions on Image Processing 17, 1047–1060 (2008)
4. Lezoray, O., Ta, V.T., Elmoataz, A.: Partial differences as tools for filtering data on graphs. Pattern Recognition Letters 31, 2201–2213 (2010)
5. Brun, L., Kropatsch, W.: Irregular pyramids with combinatorial maps. In: Amin, A., Pudil, P., Dori, D. (eds.) SPR 1998 and SSPR 1998. LNCS, vol. 1451, pp. 256–265. Springer, Heidelberg (1998)
6. Gilboa, G., Osher, S.: Nonlocal operators with applications to image processing. Multiscale Modeling and Simulation 7, 1005–1028 (2008)
7. Peyré, G., Bougleux, S., Cohen, L.: Non-local regularization of inverse problems. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 57–68. Springer, Heidelberg (2008)
8. Yun, S., Woo, H.: Linearized proximal alternating minimization algorithm for motion deblurring by nonlocal regularization. Pattern Recogn. 44 (2011)
9. Hidane, M., Lezoray, O., Ta, V., Elmoataz, A.: Nonlocal multiscale hierarchical decomposition on graphs. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 638–650. Springer, Heidelberg (2010)
10. Chambolle, A.: An algorithm for total variation minimization and applications. Journal of Mathematical Imaging and Vision 20 (2004)
11. Chan, T.F., Osher, S., Shen, J.: The digital TV filter and nonlinear denoising. IEEE Transactions on Image Processing 10, 231–241 (2001)
12. Black, M., Sapiro, G.: Edges as outliers: Anisotropic smoothing using local image statistics. In: Nielsen, M., Johansen, P., Fogh Olsen, O., Weickert, J. (eds.) Scale-Space 1999. LNCS, vol. 1682, pp. 259–270. Springer, Heidelberg (1999)
13. Nielsen, F., Nock, R.: Fast graph segmentation based on statistical aggregation phenomena. In: Ikeuchi, K. (ed.) Proceedings of the 10th IAPR Conference on Machine Vision Applications (2007)

From Points to Nodes: Inverse Graph Embedding through a Lagrangian Formulation

Francisco Escolano[1] and Edwin R. Hancock[2]

[1] University of Alicante
sco@dccia.ua.es
[2] University of York
erh@cs.york.ac.uk

Abstract. In this paper, we introduce a novel concept: *Inverse Embedding*. We formulate inverse embedding in the following terms: given a set of multi-dimensional points coming directly or indirectly from a given spectral embedding, find the mininal complexity graph (following a MDL criterion) which satisfies the embedding constraints. This means that when the inferred graph is embedded it must provide the same distribution of squared distances between the original multi-dimensional vectors. We pose the problem in terms of a Lagrangian and find that a fraction of the multipliers (the smaller ones) resulting from the deterministic annealing process provide the positions of the edges of the unknown graph. We proof the convergence of the algorithm through an analysis of the dynamics of the deterministic annealing process and test the method with some significant sample graphs.

Keywords: Graph-based Methods, Inverse Embedding, Deterministic Annealing, Lagrangian Formulation.

1 Introduction

Recently, there has been an important avenue of methods for embedding the nodes of unattributed graphs in Euclidean subspaces in a way that we assign a given number of coordinates (feature vector) to each node in the graph. Ideally, the composition of the feature vector must correlate topological distances between nodes in terms of Euclidean distances in the embedding. One of the most popular approaches for graph embedding is *spectral embedding*. The catalog of spectral methods for graph embedding includes Laplacian Eigenmaps or LEMs [1], Diffusion Maps or DMs [2], Heat Kernels or HKs [3] and Commute Times or CTs [4]. All the cited spectral embeddings rely on a function $\mathcal{F}(.)$ of the eigenvalues (encoded in a diagonal matrix Λ) and/or eigenvectors (encoded in the matrix Φ) of a proper matrix, typically a Laplacian matrix $\mathbf{L}$ of the graph or its normalized version $\mathcal{L}$. For HK and CT embeddings $\mathcal{F}(\mathbf{L}) = \Phi\mathcal{F}(\Lambda)\Phi^T = \Theta^T\Theta$, where Θ results from the Young-Householder decomposition. For CT, $\mathcal{F}(\mathbf{L}) = \sqrt{vol_X}\Lambda^{-1/2}$; for HK we have $\mathcal{F}(\mathbf{L}) = \exp\left(-\frac{1}{2}t\Lambda\right)$ where t is time; and for DT, we have $\mathcal{F}(\mathbf{L}) = \Lambda^t$ where Λ results from a generalized eigenvalue/eigenvector problem as in the case of LEM where $\mathcal{F}(\mathbf{L}) = \Phi$.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 194–201, 2011.

Alternatively to spectral methods; Bunke and co-workers [5] have proposed to embed graphs through exploiting dissimilarities (typically approximations of the graph edit distance) between a graph and a set of prototypes so that a graph is encoded by a vector of dissimilarities. The main advantage of the latter representation is that it simplifies the problem of classifying graphs [5,6] or even the problem of finding the median graph [7]. The method described herein is inspired in the way that one recovers a particular grap (the median graph) from an embedding as described in [7] where vectors encode graphs. However, in this paper, the proposed method relies on embeddings where each vector encodes a node, and what is more important, the catalog of possible embeddings relying on spectral graph theory is wider making our method more general. Helping in graph classification is also shared with spectral embedding methods when the graph is encoded by a set of feature vectors [3] (typically the columns of Θ). Our recent work in this context has been addressed to re-formulate the problem of graph matching in terms of the non-rigid alignment of two sets of feature vectors coming from different graphs [8]. In a more general context, we consider the set of feature vectors resulting from spectral graph embedding as a multi-dimensional probability distribution. Such consideration and the fact of using information-theoretic dissimilarity measures between distributions is shared with recent population-based point matching algorithms. In [8] we define the symmetrized normalized entropy squared variation (SNESV), but the catalog also includes the Henze-Penrose (KP) divergence [10] based on MSTs (Minimum Spanning Trees) and relying on the Friedman-Rafsky test [11] as well as the KD-partitions (KDP) divergence based on the method proposed by Stowell and Plumbey [12]. The latter dissimilarities have been tested both in contexts of graph comparison [8] where the SNESV is the best one, and in shape comparison where HP is the best [13].

Working with distributions of feature vectors instead of graphs opens a novel perspective, not only for classification and matching/comparison of graphs but for the building of generative models for graphs. Consider the case of a multi-dimensional distribution coming from the embedding of a graph or from the aggregation of several distributions coming from different graphs coming from the same object class. Nowadays it is possible to learn, in a very fast and consistent way, a minimal complexity Gaussian mixture model see our method proposed in [14,15]. Given the Gaussian mixture is straightforward to generate samples which will encode representations of nodes coming from the same kind of graph. What is the formal relationship between the set of samples generated and the topology of the original graphs? It seems that when we exploit embedding to map graphs to multi-dimensional distributions we loss topological information, that is, *it is difficult to come back to the original structure.* Since we want to progress in the distributional domain it is also desirable to propose methods for learning a graph from a multi-dimensional distribution. This is the motivation of *inverse embedding.* In this paper we pose the problem of inverse embedding through a Lagrangian formulation as in [16]. The formal beauty of the proposed deterministic annealing algorithm is that a fraction of the optimal Lagrange multipliers yield the MDL (minimum-description length) graph satisfying the embedding constraints. We provide a convergence proof through the analysis of the dynamics of the algorithm and test it succesfully in some representative graphs. In this paper we consider the case of CT embedding because Commute Times have proved to be more discriminative than the other spectral ones [8].

2 A Lagrangian Method for Inverse Embedding

2.1 Problem Formulation

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ be a collection of m-dimensional points in the Euclidean space and generated by a node embedding of an unknown graph $G = (V, E)$ where $|V| = m$ and adjacency matrix $\mathbf{A}$. The problem of learning or inferring the graph G from the latter collection of multi-dimensional points can be posed in the following terms:

$$\begin{aligned} Max. \quad & \sum_{j>i} A_{ij} \\ s.t. \;\; \Theta_{ij} &= ||\mathbf{x}_i - \mathbf{x}_j||^2 \\ & 0 \leq A_{ij} \leq 1; \forall\, i, j \;, \end{aligned} \tag{1}$$

where $\Theta_{ij} = ||\Theta_i - \Theta_j||^2$ and Θ_i, Θ_j are the $m-$dimensional coordinates of the embedded nodes i and j respectively. The maximization is motivated by the fact that the starting point of the method will be a complete graph which usually does not satisfies all the embedding constraints. Therefore, those links in the adjacency matrix which do not satisfy the constraints will be reduced to zero. Therefore the maximization of $\sum_{j>i} A_{ij}$ is consistent with finding the *closest graph to the complete one which satisfies all the embedding constraints*.

For the commute times embedding we commence by computing the normalized Laplacian of the graph $\mathcal{L} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$ where $\mathbf{D}$ is the diagonal degree matrix, and $\mathbf{L} = \mathbf{D} - \mathbf{A}$ is the Laplacian matrix. The *hitting time* $O(i, j)$ of a random walk on a graph is defined as the expected number of steps before node j is visited, commencing from node i. The *commute time* $CT(i, j)$, on the other hand, is the expected time for the random walk to travel from node i to reach node j and then return. As a result $CT(i, j) = O(i, j) + O(j, i)$. In terms of the Green's function the commute time is given by [4]

$$CT(i,j) = vol \sum_{z=2}^{m} \frac{1}{\lambda^{(z)}} \left(\frac{\phi^{(z)}(i)}{\sqrt{d_i}} - \frac{\phi^{(z)}(j)}{\sqrt{d_j}} \right)^2 \tag{2}$$

where $\lambda^{(z)}$ is the $z-$th eigenvalue and $\phi^{(z)}(i)$ $i-$th component of the $z-$th eigenvector of $\mathbf{L}$, $vol = trace(\mathbf{D})$ is the volume of the graph and d_i and d_j are the respective degrees of nodes i and j. Then, the CT embedding is given by the following function of the eigenvalues and eigenvectors of the normalized Laplacian $\Theta = \sqrt{vol}\Lambda^{-1/2}\Phi\mathbf{D}^{-1/2}$ where: Θ is a $m \times m$ matrix with the $i-$th column Θ_i

being the embedding coordinates of the $i-$th node, Λ is the diagonal matrix of eigenvalues ($\lambda^{(1)} = 0$ and $0^{-1/2} = 0$) and Φ is the matrix of eigenvectors. Then, the form of Θ_i is

$$\Theta_i = \sqrt{\frac{vol}{d_i}} \left(0 \ \frac{1}{\sqrt{\lambda^{(2)}}}\phi^{(2)}(i) \ldots \frac{1}{\sqrt{\lambda^{(m)}}}\phi^{m}(i) \right)^T \ . \tag{3}$$

Having an initial definition of $\Theta_{ij} = ||\Theta_i - \Theta_j||^2$ it proceeds to re-formulate the problem defined by Eq. 1 in terms of Lagrange multipliers. Therefore, using Lagrange multipliers (one for each constraint) the problem is equivalent to maximizing:

$$E(A, \{\alpha_{ij}\}) = \sum_{ij:j>i} A_{ij} + \frac{1}{\beta} \sum_{ij:j>i} A_{ij}(\log A_{ij} - 1) + \sum_{ij:j>i} \alpha_{ij}(\Theta_{ij} - ||\mathbf{x}_i - \mathbf{x}_j||^2) \ , \tag{4}$$

where the second (entropic) term is used to make concave the energy function for lower values of β; the third term contains the $m(m+1)/2 - m$ Lagrange multipliers (one multiplier per constraint).

The fixed point equations for updating the A_{ij} are given by

$$\begin{aligned} \frac{\partial E}{\partial A_{ij}} &= 1 + \frac{1}{\beta}\log A_{ij} + \alpha_{ij}\frac{\partial \Theta_{ij}}{\partial A_{ij}} \\ \frac{\partial E}{\partial A_{ij}} &= 0 \ \Rightarrow \ \frac{1}{\beta}\log A_{ij} = -1 - \alpha_{ij}\frac{\partial \Theta_{ij}}{\partial A_{ij}} \\ &\Rightarrow \quad A_{ij} = \exp \beta \left(-1 - \alpha_{ij}\frac{\partial \Theta_{ij}}{\partial A_{ij}} \right) \ , \end{aligned} \tag{5}$$

where $\frac{\partial \Theta_{ij}}{\partial A_{ij}}$ (approximated numerically) is the gain in terms of squared distance with respect to the variation of a single component A_{ij}.

On the other hand, the update of the multipliers has not a closed formed and it must be performed through gradient ascent, given the previous multipliers and distances:

$$\frac{\partial E}{\partial \alpha_{ij}} = \Theta_{ij} - ||\mathbf{x}_i - \mathbf{x}_j||^2 \ \Rightarrow \ \alpha_{ij}^{t+1} = \alpha_{ij}^t + \mu(\Theta_{ij}^t - ||\mathbf{x}_i - \mathbf{x}_j||^2) \ , \tag{6}$$

with $\mu \in [0, 1]$ (learning factor). In practice, such a factor must be fixed so that it decreases with the size of the graph.

2.2 Deterministic Annealing Algorithm

Given the above updates for $\{A_{ij}\}$ and $\{\alpha_{ij}\}$ it is straightforward to devise a deterministic annealing algorithm:

```
Initialize $\beta$ to $\beta_0$, $A_{ij} = 1/m, \alpha_{ij} = 0, j > i, \mu$
Begin: Deterministic Annealing. Do while $\beta \leq \beta_f$
    $H \leftarrow ComposeAdjacencyMatrix(\{A_{ij}\})$
    $\Theta \leftarrow Embedding(H)$
    $\alpha_{ij} \leftarrow \alpha_{ij} + \mu(\Theta_{ij} - ||\mathbf{x}_i - \mathbf{x}_j||^2)$
    $\frac{\partial \Theta_{ij}}{\partial A_{ij}} \leftarrow ComputeDerivative(i, j, G)$
    $A_{ij} \leftarrow \exp \beta \left(-1 - \alpha_{ij} \frac{\partial \Theta_{ij}}{\partial A_{ij}}\right)$
    $\beta \leftarrow \beta\beta_r$
End
$G = MDLCleanup(\{A_{ij}\}, \{\alpha_{ij}\})$
```

Convergence Proof. In the latter algorithm, the initialization $A_{ij} = 1/m$, that is, a barycenter depending on the complete graph ($A_{ij} = 1$) ensures that the $m-$dimensional points of Θ obtained from *Embedding*(G) are equally spaced. In practice this implies large equal square distances Θ_{ij}. We will have both $\Theta_{ij} - ||\mathbf{x}_i - \mathbf{x}_j||^2 < 0$ and $\Theta_{ij} - ||\mathbf{x}_i - \mathbf{x}_j||^2 > 0$. However, as the unknown graph can only be obtained by removing edges from the complete graph (that is zeroing components of the barycenter adjacency matrix) the cases where $\Theta_{ij} - ||\mathbf{x}_i - \mathbf{x}_j||^2 < 0$ will start dominating over the others. Consequently, as many $\alpha_{ij} < 0$ we wil have that $\sum_{ij:j>i} \alpha_{ij}(\Theta_{ij} - ||\mathbf{x}_i - \mathbf{x}_j||^2) > 0$. It is straightforward to prove that under the conditions of $A_{ij} < 0$ we will obtain $\frac{\partial \Theta_{ij}}{\partial A_{ij}} > 0$. Therefore, at the beginning of the determinisic annealing we will have $A_{ij} = \exp \beta(-1 + z)$, $z > 0$, where the magnitude of z depends on μ. As the effect of the exponentiation for a low β is to make equall all the A_{ij} unless significant differences appear, we will obtain $A_{ij}^{t+1} = kA_{ij}^t, k \in (0, 1)$ for low and mid values of β (while the $\frac{1}{\beta} \sum_{ij:j>i} A_{ij}(\log A_{ij} - 1) < 0$) dominates the energy. This latter term, which also plays the role of a regularizing term for the A_{ij}, will increase with β more significantly than $\sum_{ij:j>i} \alpha_{ij}(\Theta_{ij} - ||\mathbf{x}_i - \mathbf{x}_j||^2) > 0$.

Considering the update of the multipliers α_{ij}, the latter entropic term has a significant effect on $\alpha_{ij}^{t+1} = \alpha_{ij}^t + \mu(\Theta_{ij}^t - ||\mathbf{x}_i - \mathbf{x}_j||^2)$ which is focused on the values of Θ_{ij}^t. We have that since $A_{ij}^{t+1} < A_{ij}^t$ for small β, we obtain $\Theta_{ij}^{t+1} < \Theta_{ij}^t$ in these conditions. However, a good property of the proposed algorithm is that despite the latter effect, the entropic term is unable to regularize the α_{ij} where each multiplier evolves in order to satisfy its constraint. As β increases, we have that small variations in α_{ij} yield also small variations in $A_{ij} = \exp \beta(-1 - \alpha_{ij} \frac{\partial \Theta_{ij}}{\partial A_{ij}})$. There are two consequences: (i) some multipliers α_{ij} tend to zero faster than others, which indicates a greater degree of constraints satisfaction; and (ii) the corresponding A_{ij} tend also to zero faster than others, which indicates that the structure of the hidden graph is emerging: while maximizing the sum of A_{ij} constrained to satisfying the embedding constraints we obtain close-to-zero values in positions where constraints are satisfied and this is enforced by the exponential decay. When β_f is reached, the smalles values of A_{ij} and the less negative α_{ij} yield the solution to the inverse embedding problem.

Regarding the function $G = MDLCleanup(\{A_{ij}\}, \{\alpha_{ij}\})$ it must be carefully designed for inferring the correct graph. For instance, as the number of edges in the graph is typically very low compared with the one of a complete graph, a clustering method working with the A_{ij}, with α_{ij} or with their combination will fail due to the fact that we have a very small cluster and a large. The usually tiny numerical differences between the correct edge-weights/multipliers and the incorrect ones prevent the use of a generic threshold. Then, we propose to apply a MDL (minimum description length criterion) consisting on sorting the weights/multipliers in ascending order and taking the first n ones (taking the absolute values in the case of the multipliers) leading to a graph with a unique connected component. Considering that the minimum mumber of edges for connecting m nodes is $m-1$ we set $n = m-1$, check if the graph is connected and otherwise we increase n until connectivity arises. Therefore, *MDL-cleanup provides the smallest connected graph satisfying the embedding constraints.* In practice it is more convenient to use the multipliers because they converge faster than the edge weights which become more ambiguous. In some sense we are applying a kind of *explicit mild MDL* because we do not select the correct model (graph) order until the algorithm converges. However, MDL is implicit in the computation of the multipliers and in their ranking; we must only find the correct cut-off.

In Fig. 1 we represent data concerning both a Delaunay triangulation and a linear graph. The energy functions plots show the convergence of the algorithm. On the other hand, the plot of sorted absolute values of the optimal Lagrange multipliers for the Delaunay triangulation shows also the MDL cutoff found by the algorithm (slightly greater than the corresponding to the ground truth). In terms of error, if we measure the relative error $\epsilon = \sum_{ij} \frac{|G_{ij} - G^*_{ij}|}{|G|^2}$ which is the relative number of different components in the adjacency matrix we obtain

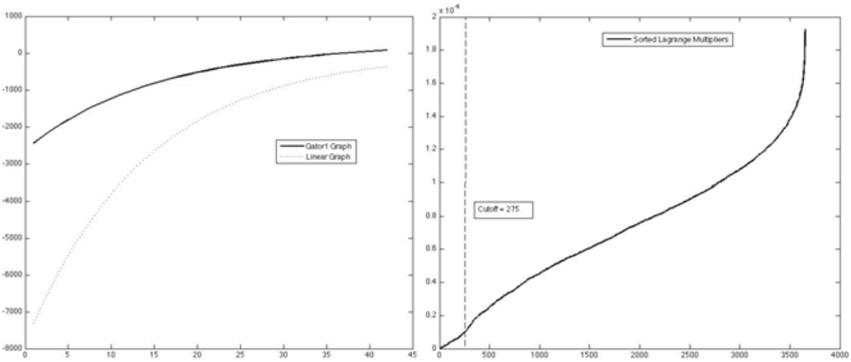


Fig. 1. Examples of applications of the deterministic annealing algorithm. Left: Evolution of the energy function for the Delaunay triangulation (first graph of the GatorBait database) of $m = 86$ nodes and for the linear graph of $m = 50$ nodes. Right: Sorted optimal Lagrange multipliers for the Delaunay triangulation with the obtained MDL cutoff ($k = 275$ edges); in this case the real number of edges is 242.

$\epsilon = 0.0224$ (2.24%) for the Delaunay triangulation and $\epsilon = 0.0$ for the linear graph.

3 Conclusions

In this paper we have formulated the problem of inverse embedding and have proposed a simple method for solving it. We show that the solution to this inference problem is typically encoded by a small fraction of the optimal Lagrange multipliers. Such small fraction is key for defining a MDL-like cleanup strategy which returns the connected graph which satisfies the embedding constraints with the minimal number of edges. Future work includes a more in depth test with complete graph databases in order to evaluate the method more thoroughly.

Acknowledgements. F. Escolano is supported by the project TIN2008-04416 of the Spanish Government. E. R. Hancock is supported by the EU FET project SIMBAD and a Royal Society Wolfson Research Merit Award.

References

1. Belkin, M., Niyogi, P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Computation 15(6), 1373–1396 (2003)
2. Nadler, B. Lafon, L., Coifman, R., Kevrekidis, I.G.: Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck Operators. In: Proc. of NIPS 2005 (2005)
3. Xiao, B., Hancock, E.R., Wilson, R.C.: Geometric Characterization and Clustering of Graphs using Heat Kernel embeddings. Image Vision Comput. 28(6), 1003–1021 (2010)
4. Qiu, H., Hancock, E.R.: Clustering and Embedding Using Commute Times. IEEE Trans. on PAMI 29(11), 1873–1890 (2007)
5. Bunke, H., Riesen, K.: Graph Classification Based on Dissimilarity Space Embedding. In: da Vitoria Lobo, N., Kasparis, T., Roli, F., Kwok, J.T., Georgiopoulos, M., Anagnostopoulos, G.C., Loog, M. (eds.) S+SSPR 2008. LNCS, vol. 5342, pp. 996–1007. Springer, Heidelberg (2008)
6. Riesen, K., Bunke, H.: Feature Ranking Algorithms for Improving Classification of Vector Space Embedded Graphs. In: Jiang, X., Petkov, N. (eds.) CAIP 2009. LNCS, vol. 5702, pp. 377–384. Springer, Heidelberg (2009)
7. Ferrer, M., Valveny, E., Serratosa, F., Riesen, K., Bunke, H.: Generalized Median Graph Computation by Means of Graph embedding in Vector spaces. Pattern Recognition 43(4), 1642–1655 (2010)
8. Escolano, F., Hancock, E.R., Lozano, M.: Graph Matching through Entropic Manifold Alignment. In: Proc. of CVPR 2011 (2011) (accepted for publication)
9. Chen, T., Vemuri, B., Rangarajan, A., Eisenschenk, S.: Group-wise Point-set Registration using a Novel cdf-based Havrda-Charvát Divergence. International Journal of Computer Vision 86(1), 111–124 (2010)
10. Henze, N., Penrose, M.: On the multi-variate runs test. Annals of Statistics 27, 290–298 (1999)

11. Friedman, J.H., Rafsky, L.C.: Mutivariate Generalization of the Wald-Wolfowitz and Smirnov Two-Sample Tests. Annals of Statistics 7(4), 697–717 (1979)
12. Stowell, D., Plumbley, M.D.: Fast Multidimensional Entropy Estimation by K-d Partitioning. IEEE Signal Processing Letters 16(6), 537–540 (2009)
13. Escolano, F., Lozano, M.A., Bonev, B., Suau, P.: Bypass Information-Theoretic Shape Similarity from Non-rigid Points-based Alignment. In: Proc. of CVPR Workshops (NORDIA) (2010)
14. Peñalver, A., Escolano, F., Sáez, J.M.: Learning Gaussian Mixture Models With Entropy-Based Criteria. IEEE Transactions on Neural Networks 20(11), 1756–1771 (2009)
15. Peñalver, A., Escolano, F., Bonev, B.: Entropy-Based Variational Scheme for Fast Bayes Learning of Gaussian Mixtures. In: Proc. of S+SSPR 2010 (2010)
16. Rangarajan, A., Yuille, A.L., Mjolsness, E.: Convergence Properties of the Softassign Quadratic Assignment Algorithm. Neural Computation 11(6), 1455–1474 (1999)

K-nn Queries in Graph Databases Using M-Trees[*]

Francesc Serratosa, Albert Solé-Ribalta, and Xavier Cortés

Universitat Rovira i Virgili, Departament d'Enginyeria Informàtica i Matemàtiques, Spain
{francesc.serratosa,albert.sole}@urv.cat,
xavier.cortes@estudiants.urv.cat

Abstract. Metric trees (m-trees) are used to organize and execute fast queries on large databases. In classical schemes based on m-trees, routing information kept in an m-tree node includes a representative or a prototype to describe the sub-cluster. Several research has been done to apply m-trees to databases of attributed graphs. In these works routing elements are selected graphs of the sub-clusters. In the current paper, we propose to use Graph Metric Trees to improve k-nn queries. We present two types of Graph Metric Trees. The first uses a representative (Set Median Graph) as routing information; the second uses a graph prototype. Experimental validation shows that it is possible to improve k-nn queries using m-trees when noise between graphs of the same class is of reasonable level.

Keywords: graph database, m-tree, graph organization, graph prototype, graph indexing, Mean Graph, Median Graph, Set Median Graph.

1 Introduction and Related Work

Indexing structures are fundamental tools in database technology. In the field of pattern recognition, they are used to obtain efficient access to large collections of images. Traditional database systems manage global properties of images, such as histograms, and many techniques for indexing multi-dimensional data sets have been defined. Since a distance function over a particular attribute domain always exists, this distance function can be used to partition the data. In addition, it can be exploited to efficiently support queries. Several multi-dimensional indexes have been developed, such as, color, texture, shape and so on, with the aim of increasing the efficiency in executing queries on sets of objects characterized by multi-dimensional features [1]. Effective access to image databases requires queries addressing the expected appearance of searched images [2]. To overcome these systems, objects contained in images can be modeled using attributed graphs, [3] and [4]. In this way, each object is composed by local entities and relations between these entities. This paradigm enriches indexing structures considering of the same relevance local entities and relations. Likewise traditional indexing structures, each local entity can represent multi-dimensional features. The main impediment of using attributed graphs as

[*] This research is supported by Consolider Ingenio 2010: project CSD2007-00018 & by the CICYT project DPI 2010-17112.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 202–210, 2011.

indexing structures is that the problem of optimally compare two attributed graphs is NP-Complete [5]. However, some sub-optimal solutions exist [6], [7].

Some indexing techniques have been developed for graph databases. We divide these techniques into two categories. In the first ones, the index is based on several tables and filters [8], [9]. In the second ones, the index structure is based on trees [3], [10], [11]. In the first group of techniques, we emphasize the method developed by Shasha *et. al.* [8] called GraphGrep. GraphGrep is based on a table in which each row stands for a path inside the graph (up to a threshold length) and each column stands for a graph. Each entry in the table is the number of occurrences of the path in the graph. More recently, Yan *et. al.* [9] proposed GIndex which uses frequent patterns as indexing features. These frequent patterns reduce the indexing space as well as improve the filtering rate. The main drawback of these models is that the construction of the indices requires an exhaustive enumeration of the paths or fragments which increases memory and time requirements. Considering the second group, the first work where metric trees were applied to graph databases was done by Berretti *et. al.* [3]. AGs were clustered hierarchically according to their mutual distances and indexed by m-trees [12]. The method in [3] was successfully applied to the problem of performing similarity queries. Latter, Lee *et. al.* [10] used this technique to model graphical representations of foreground and background scenes in videos. More recently, He and Singh [11] proposed what they called a Closure-tree. It uses a similar structure than the one presented by Berretti [3] but, the representative of the cluster was not one of the graphs but a graph prototype (called Closure Graph) which could be seen as the union of the AG that compose the cluster.

Our proposal is to apply metric trees to the problem of graph k-nn queries. In [19], metric trees where used to obtain graphs sorted by minimum distance. We propose to use as a routing element of the metric tree two types of graphs: a representative (Set Median Graph) and a graph prototype. Results validate that under a reasonable noise level graph metric trees can perform k-nn queries faster than the traditional method.

The article is structured as follows. Section 2 gives some basic definitions. Sections 3 and 4 present the metric tree and how it is applied to graph k-nn queries. In Section 5, we experimentally evaluate the models. We finish the paper drawing some conclusions and presenting the future work.

2 Graph Preliminaries

An **Attributed Graph** *AG* over *(Δ_v* and *Δ_e)* is defined by a tuple $G = (\Sigma_v, \Sigma_e, \gamma_v, \gamma_e)$, where $\Sigma_v = \{v_k \mid k = 1, \ldots, R\}$ is the set of vertices, $\Sigma_e \in \{e_{ij} \mid i, j \in \{1, \ldots, R\}, i \neq j\}$ is the set of arcs and $\gamma_v : \Sigma_v \rightarrow \Delta_v$ and $\gamma_e : \Sigma_e \rightarrow \Delta_e$ assign attribute values to vertices and arcs respectively. In case it is required, any AG can be extended with null nodes, which have special attribute $\emptyset \in \Delta_v$ [13].

Set Median Graph. Given a set of graphs $\Gamma = \{G^1, G^2, \ldots, G^N\}$ and a distance between graphs $d(G^i, G^j)$, the *Set Median Graph* is defined as follows.

$$\bar{g} = \underset{g \in \Gamma}{argmin} \sum_{G^i \in \Gamma} d(g, G^i) \tag{1}$$

That is, the graph in the set Γ that minimizes distances to all other graphs in Γ. For further references to the Set Median Graph the reader is referred to [16].

Mean Graph: Given a set of graphs $\Gamma = \{G^1, G^2, \ldots, G^N\}$, we define a common labeling [15] $\psi = \{h^1, h^2, \ldots, h^n\}$ over the graphs in Γ as a set bijective mappings from nodes of graphs in Γ to a virtual vertex set L_v , $h^i = \Sigma_v^i \rightarrow L_v$. We define the Mean Graph from a set of attributed graphs Γ under a common labeling ψ as another attributed graph where attributes on nodes and arcs are:

$$\gamma_v(v_j) = \frac{1}{M} \sum_{\forall G^i \in \Gamma | \gamma_v\left(h^{i^{-1}}(v_j)\right) \neq \phi} \gamma_v\left(h^{i^{-1}}(v_j)\right) \text{ and } \gamma_e(e_{kj}) = \frac{1}{N} \sum_{\forall G^i \in \Gamma | \varepsilon \neq \phi} \varepsilon \tag{2}$$

being M the number of nodes such that $\forall G^i \in \Gamma | \gamma_v\left(h^{i^{-1}}(v_j)\right) \neq \phi$, N the number of arcs such that $\varepsilon \neq \phi$ and being ε the attribute value of the arc $e_{h^{i^{-1}}(v_k), h^{i^{-1}}(v_j)}$ in G^i.

3 K-nn Queries Based on m-Trees

A metric tree [12], m-tree in short, is a tree of nodes, each containing a fixed maximum number of m entries, $< node > := \{< entry >\}^m$. In turn, each entry is constituted by a routing element H, a reference to the root r^H of a sub-index containing the element in the so-called covering region of H and a radius d^H providing an upper bound for the distance between H and any element in its covering region, $< entry > := \{H, r^H, d^H\}$.

To perform k-nn queries in metric trees, the tree is analyzed in a top down fashion using triangular inequality to prune unfruitful tree branches. The method proposed in [12] uses mainly two arrays PR and NN. Array PR stores the possibly fruitful tree nodes to be explored and NN stores the best K graphs found until the moment.

Let G be a query graph and d_k the maximum distance from G to any element in NN. In each iteration of the search algorithm, the tree node in PR with lower distance to G is selected, let this node be named TN_f. Children of $TN_f = TN_f^{1..last}$ are analyzed and its distance to G is computed. If son i is a routing node, this node is inserted in PR if $d(TN_f^i, G) - r^{TN_f^i} \leq d_k$. In other words, it is possible to find a graph with lower distance than the ones already found. If son i is a leave, that is, a database graphs, and $d(TN_f^i, G) \leq d_k$, array NN and d_k are updated to consider this element. Note that in k-nn queries, d_k acts as a dynamic maximum search distance of range queries. For a detailed description of k-nn and range queries the reader is referred to the original article [12].

4 Graph Metric Trees

In this section, we first present the qualities of Set Median Graph and Mean Graphs as routing elements. Second, we present a method to obtain a metric tree where routing nodes are Set Median Graphs or Mean Graphs.

4.1 Graph Metric Tree Routing Nodes

As it was described in Section 2, each non leave m-tree node is composed, among others, of a routing element H. Classical approaches use m-tree to organize numerical data. When the format of the data is a numerical vector of features, it is easy to choose between a representative of the sub-cluster or a prototype of it. Nevertheless, when the format of the data is a graph, it is more difficult to decide whether it better to represent the sub-cluster by a representative or a prototype. The main advantage of using a graph representative, such as a Set Median Graph, is that its computation is straightforward and fast. However, it is statistically difficult to find a good representative if the number of elements in the set is small. See Fig. 1.1. On the other side, a graph prototype offers a high quality representation of the set, which should improve performance with respect to use a representative. But it is computationally hard to construct it. See Fig. 1.2. Applied to the specific problem of constructing a Graph Metric Tree, the main effect of using a graph prototype is the reduction of the overlap between sub-clusters, due to the radius of the covering region can be more tightly adjusted. In fact, if we use graph prototype as a routing element, the radius of the covering region has to be equal or lower than the radius of the covering region represented by one of the AGs of the set.

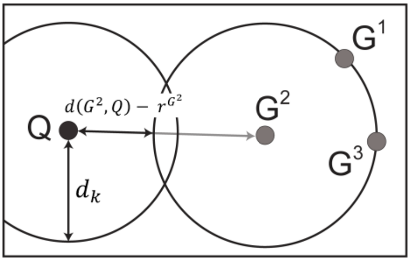


Fig. 1.1. Clusters represented by an AG

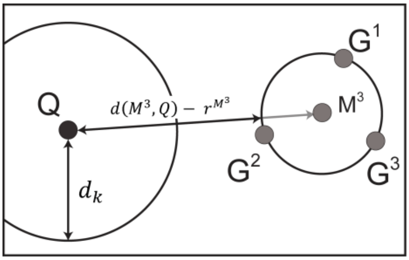


Fig. 1.2. Clusters represented by a computed graph prototype

Fig. 1.1 and 1.2 shows an example where the same query is performed using the two types of routing nodes. In the example, Q is compared to a graph cluster that represents graphs G^1, G^2 and G^3. In Fig. 1.1 the cluster is represented by a representative and in Fig. 1.2 the cluster is represented by a graph prototype. In the given example the cluster represented in Fig. 1.1 must be explored due to $d(G^2,G) - r^{G^2} \leq d_k$. However, in Fig. 1.2 the cluster radius is better adjusted and we can ensure that it will not have any desired graph due to $d(M^3,G) - r^{M^3} > d_k$. Note that, in metric trees, the smaller the radius of clusters the lower the number of comparisons that we must perform to find the solution.

4.2 Computation of a Graph Metric Tree

We provide a construction methodology, based on clustering techniques, from which we are able to construct exactly the same metric tree independently of the type and performance of the routing nodes. We use a non-balanced tree constructed through a hierarchical clustering [17] algorithm and *"average"* linkage clustering. In this way, given a set of graphs, we first compute the pairwise distance matrix over the whole set of graphs and then we construct a dendogram. We obtain a set of partitions that clusters the set of AGs using some horizontal cuts over the dendogram. With the resulting partitions, we generate the Graph Metric Tree and we compute the routing nodes. Fig. 2.1 shows an example of a dendogram. The AGs G^i are placed on the leaves of the dendogram and the routing elements M^j are placed on the junctions between the cuts and the horizontal lines of the dendograms. Fig. 2.2 shows the obtained m-tree.

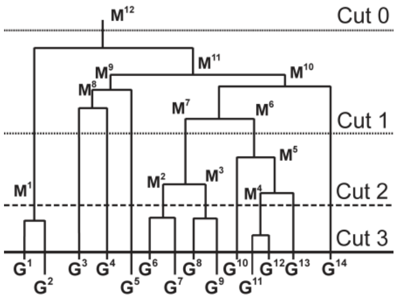


Fig. 2.1. Example of a dendogram using 14 graphs

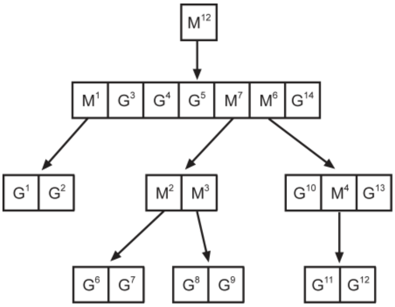


Fig. 2.2. Metric tree obtained by dendogram in Fig. 2.1

Computing the Graph Metric Tree Based on a Representative of the Sub-set

The idea of using a graph representative to compute a graph metric tree was first presented in [3] by Berretti. Here, we adapt [3] ideas to construct the m-tree based on Set Median Graphs. Given a pairwise distance matrix of the whole set of AGs, the computation of the representative AG, given a sub-set, is simply performed by adding the pre-computed distances between the involved AGs. For instance, to obtain the M^7 that appears at Fig. 2.1, we use the distances between the AGs G^6, G^7, G^8 and G^9.

The covering region radius r^p of the AG M^p is taken as the maximum distance between M^p and any of the AGs in the sub-set.

Computing the Mean Graph Metric Tree

To compute the Mean Graph m-tree, we first need to compute the common labeling between the graphs of the cluster, to this aim we propose to use the Graduated Assignment Common Labeling algorithm presented in [15]. Once the common labeling is computed each Mean Graph is computed using (2).

The covering region radius r^p of the Mean Graph M^p is computed applying three rules, depending whether the type of the descendant of M^p in the dendogram is another Mean Graph (that is, a routing node of the m-tree) or an AG (that is, a leaf of the m-tree):

- When both descendants are AGs (G^a and G^b):

$$r^p = max\left(Dist(M^p, G^a), Dist(M^p, G^b)\right) \tag{3}$$

- When a descendant is a Mean Graph (M^a) and the other is an AG (G^b):

$$r^p = max\left(Dist(M^p, M^a) + r^a, Dist(M^p, G^b)\right) \tag{4}$$

- When both descendants are Mean Graphs (M^a and M^b):

$$r^p = max(Dist(M^p, M^a) + r^a, Dist(M^p, M^b) + r^b) \tag{5}$$

Second (4) and third (5) rule are illustrated in Fig. 3.1 and 3.2 respectively.

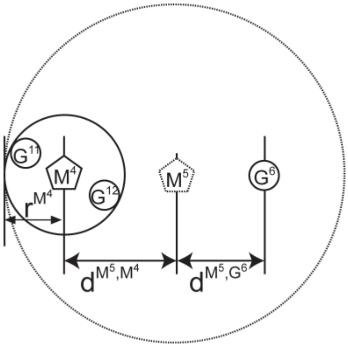


Fig. 3.1. Second rule to compute the covering region radius

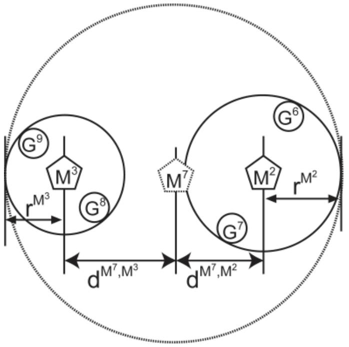


Fig. 3.2. Third rule to compute the covering region radius

5 Evaluation

To evaluate the performance of graph m-trees we apply them to the problem of classification. We used two indices to evaluate both models: the penetration rate and the classification rate.

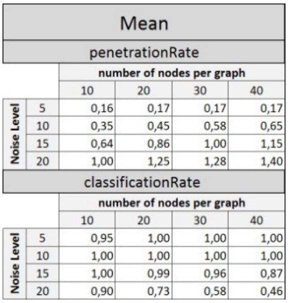
Penetration Rate. The index is addressed to evaluate the capacity of the m-tree to properly route k-nn the queries [18]. This index evaluates the quantity of nodes accessed in comparison to the number of graphs used to construct the m-tree:

$$Penetration_{rate} = \frac{number\ of\ accessed\ nodes}{number\ of\ graphs\ used\ in\ construction} \tag{6}$$

Note that the m-tree contains more graphs than the graphs used in construction due to some attributed graphs are introduced for routing purposes.

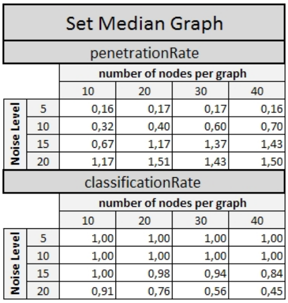
Classification Performance Rate. In the ideal case, the k-nn search applied to any m-tree should return the same information than a classical k-nn search. However, recall that Graph Edit Distance does not hold the triangle inequality, so routing over graph metric trees using triangle inequality theorem could produce different and non-correct results. Taking into account this consideration, we consider necessary to evaluate the classification rate of graph metric trees.

To evaluate the graph metric tree proposed here under controlled graph order and known noise level, we created synthetically $16 * 5 = 80$ tests. Each test is composed by 15 classes of $N = [10 + Q]$ graphs per class, each graph of order $or = [10, 20, 30, 40]$. Each class was created as follows. We randomly generate a base graph with random attributes in the range Δ_v*=[0..100, 0..100]*. Edges are defined by the Delaunay triangulation. Then, with this base graph, we created the *N* class graphs by: 1) generating Gaussian noise at every node with standard deviation *σ={0.05, 0.1, 0.15, 0.20}*, 2) removing *v ∈ [5%, 10%, 15%, 20%]* nodes randomly, 3) inserting *v* nodes (with random attributes) and 4) changing the state of *v* edges. We created a single Graph Metric Tree per test. To do so, the distance matrix was obtained using the Graph Edit Cost defined in [14]. The bijection which leads to the minimum cost was computed using the Graduated Assignment algorithm [6]. We took values of $K_n = 10, K_e = 10$. Using the $Q = 7$ resting graphs, we performed Q 3-nn queries per class. That is, a total of $105 * 5 = 525$ queries per test. For each test, we evaluated the penetration rate and the classification performance. With the aim of obtaining non-biased results, we performed 5 experiments per each test and we averaged the results.



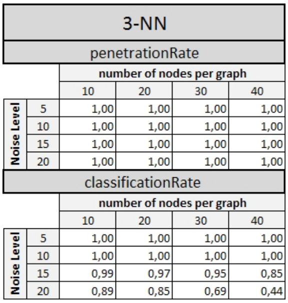
Mean					
penetrationRate					
		number of nodes per graph			
		10	20	30	40
Noise Level	5	0,16	0,17	0,17	0,17
	10	0,35	0,45	0,58	0,65
	15	0,64	0,86	1,00	1,15
	20	1,00	1,25	1,28	1,40
classificationRate					
		number of nodes per graph			
		10	20	30	40
Noise Level	5	0,95	1,00	1,00	1,00
	10	1,00	1,00	1,00	1,00
	15	1,00	0,99	0,96	0,87
	20	0,90	0,73	0,58	0,46

Fig. 4.1 Results using Mean Graph



Set Median Graph					
penetrationRate					
		number of nodes per graph			
		10	20	30	40
Noise Level	5	0,16	0,17	0,17	0,16
	10	0,32	0,40	0,60	0,70
	15	0,67	1,17	1,37	1,43
	20	1,17	1,51	1,43	1,50
classificationRate					
		number of nodes per graph			
		10	20	30	40
Noise Level	5	1,00	1,00	1,00	1,00
	10	1,00	1,00	1,00	1,00
	15	1,00	0,98	0,94	0,84
	20	0,91	0,76	0,56	0,45

Fig. 4.2 Results using the Set Median Graph



3-NN					
penetrationRate					
		number of nodes per graph			
		10	20	30	40
Noise Level	5	1,00	1,00	1,00	1,00
	10	1,00	1,00	1,00	1,00
	15	1,00	1,00	1,00	1,00
	20	1,00	1,00	1,00	1,00
classificationRate					
		number of nodes per graph			
		10	20	30	40
Noise Level	5	1,00	1,00	1,00	1,00
	10	1,00	1,00	1,00	1,00
	15	0,99	0,97	0,95	0,85
	20	0,89	0,85	0,69	0,44

Fig. 4.3. Ground truth using 3-nn

Fig. 4.1, 4.2 and 4.3 show the results of the evaluation. Analyzing metric trees classification rate in comparison with the ground truth 3-nn we see that the performance is not affected. That is, we conclude that, at least in the used dataset, the Graph Edit Distance fulfils metric properties.

With respect to penetration rate we could say that there is a noise level range where graph metric trees are faster than traditional k-nn queries. However, the penetration rate of metric trees decays when noise exceeds level 15. Penetration rate is also affected with the number of nodes of the graph, being lower with graph of small order. In addition, we see that with values of noise higher than 15, the penetration rates crosses the critical rate of 1 meaning that it is faster to use a classical 3-nn classifier than a 3-nn classifier using graph metric trees.

Comparing the results obtained using a graph representative or a graph prototype, we conclude that using a graph prototype do not significantly affect neither the penetration rate nor the classification rate.

6 Conclusions and Further Work

In this article, we presented an approach to increment performance in k-nn graph queries. This approach makes use of metric trees adapted to attributed graph data. We tested the approach using two types of graph routing elements: a representative (Set Median Graph) and a graph prototype, which we call Mean Graph, proposed in the present article.

Results show that metric trees can be used to improve k-nn graph queries when the noise among graphs that belong to the same class is not excessively high. Moreover, Mean Graph Metric Tree seems to be more effective in comparison to Set Median Metric Tree but the improvement is not significant.

Besides the direct analysis of the results, it is worth to note that under some conditions, metric properties, such as triangle inequality, can be used on Graph Edit Distance; even this distance is not proven to be a metric.

As a further work, authors will attempt to deduce under which conditions Edit Distance behaves as a metric and how this metric properties differ from the ideal case.

References

1. Smith, J.R., Samet, H.: VisualSEEk: A Fully Automated Content-Based Image Query System. In: Proc. ACM Multimedia, pp. 87–98 (1996)
2. Gudivada, V.N., Raghavan, V.V.: Special issue on Content Based Image Retrieval Systems. Computer 28(9) (1995)
3. Berretti, S., Del Bimbo, A., Vicario, E.: Efficient Matching and Indexing of Graph Models in Content-Based Retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(10), 1089–1105 (2001)
4. Zhao, J.L., Cheng, H.K.: Graph Indexing for Spatial Data Traversal in Road Map Databases. Computers & Operations Research 28, 223–241 (2001)
5. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness (1979)
6. Gold, S., Rangarajan, A.: A Graduated Assignment Algorithm for Graph Matching. IEEE Transactions on Pattern Analysis and Machine Intelligence 18(4), 377–388 (1996)
7. Riesen, K., Bunke, H.: Approximate graph edit distance computation by means of bipartite graph matching. Image Vision Comput. 27(7), 950–959 (2009)
8. Shasha, D., Wang, J.T.L., Giugno, R.: Algorithmics and applications of tree and graph searching. In: ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 39–52 (2002)
9. Yan, X., Yu, P.S., Han, J.: Graph indexing: a frequent structure-based approach. In: ACM SIGMOD International Conference on Management of Data, pp. 335–346 (2004)

10. Lee, S.Y., Hsu, F.: Spatial Reasoning and Similarity Retrieval of Images using 2D C-Strings Knowledge Representation. Pattern Recognition 25(3), 305–318 (1992)
11. He, H., Singh, A.K.: Closure-Tree: An Index Structure for Graph Queries. In: Proc. International Conference on Data Engineering, p. 38 (2006)
12. Ciaccia, P., Patella, M., Zezula, P.: M-tree: An Efficient Access Method for Similarity Search in Metric Spaces. In: Proc. 23rd VLDB Conference, pp. 426–435 (1997)
13. Wong, A.K.C., et al.: Entropy and distance of random graphs with application to structural pattern recognition. IEEE Trans. on Patt. Anal. & Machine Intelligence 7, 599–609 (1985)
14. Sanfeliu, A., King-Sun, F.: A Distance measure between attributed relational graphs for pattern recognition. IEEE Trans. on Systems, Man, and Cybernetics 13(3), 353–362 (1983)
15. Solé-Ribalta, A., Serratosa, F.: Graduated Assignment Algorithm for Finding the Common Labelling of a set of Graphs. In: Proceedings of Syn. and Struc. Patt. Recog., pp. 180–190 (2010)
16. Ferrer, M., Valveny, E., Serratosa, F.: Median graphs: A genetic approach based on new theoretical properties. Pattern Recognition 42(9), 2003–2012 (2009)
17. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, 2nd edn. Springer, New York (2009) ISBN 0-387-84857-6
18. Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: Handbook of Fingerprint Recognition. Springer, Heidelberg (2005) ISBN-13: 978-0387954318
19. Solé, A., Serratosa, F., Vidiella, E.: Graph Indexing and Retrieval based on Median Graphs. In: Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., Kittler, J. (eds.) MCPR 2010. LNCS, vol. 6256, pp. 311–321. Springer, Heidelberg (2010)

User-Steered Image Segmentation Using Live Markers

Thiago Vallin Spina, Alexandre Xavier Falcão, and Paulo André Vechiatto Miranda

Institute of Computing – University of Campinas (UNICAMP),
13082-852, Campinas, SP, Brazil

Abstract. Interactive image segmentation methods have been proposed based on region constraints (user-drawn markers) and boundary constraints (anchor points). However, they have complementary strengths and weaknesses, which can be addressed to further reduce user involvement. We achieve this goal by combining two popular methods in the *Image Foresting Transform* (IFT) framework, the differential IFT with optimum seed competition (DIFT-SC) and live-wire-on-the-fly (LWOF), resulting in a new method called *Live Markers* (LM). DIFT-SC can cope with complex object silhouettes, but presents a leaking problem on weaker parts of the boundary. LWOF provides smoother segmentations and blocks the DIFT-SC leaking, but requires more user interaction. LM combines their strengths and eliminates their weaknesses at the same time, by transforming optimum boundary segments from LWOF into internal and external markers for DIFT-SC. This hybrid approach allows linear-time execution in the first interaction and sublinear-time corrections in the subsequent ones. We demonstrate its ability to reduce user involvement with respect to LWOF and DIFT-SC using several natural and medical images.

1 Introduction

Image segmentation requires *object recognition* to indicate its whereabouts in the image and make corrections, and *object delineation* to define its precise spatial extent in the image. Humans usually outperform computers in recognition and the other way around can be observed in delineation. Besides, interactive segmentation is necessary in many applications, such as medical image analysis and digital matting. Hence, it is desirable to have interactive methods that combine the superior abilities of humans for recognition with the outperformance of computers for delineation in a synergistic way [1,2,3,4,5,6,7]. In this context, however, the challenges ought to simultaneously (i) maximize accuracy, precision, and computational efficiency, (ii) minimize user involvement and time, and (iii) maximize the user's control over the segmentation process.

Many interactive methods exploit boundary constraints, such as anchor points, or region constraints, such as internal and external markers, and make direct/indirect use of some image-graph concept, such as arc weight between pixels. The weight may represent different attribute functionals such as similarity, speed function, affinity, cost, distance, etc; depending on different frameworks used, such as watershed, level sets, fuzzy connectedness, graph cuts, etc [6]. In the first case, the object may be defined by optimum boundary segments that pass through the anchor points to close its boundary. This idea was first formulated as a heuristic search problem in an image-graph by

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 211–218, 2011.

Martelli [8], but with no guarantee of success. This guarantee was only possible without any shape constraints in the 2D dynamic programming framework of *live wire* [1,2]. However, the real-time response of live wire with respect to user's actions strongly depended on the image size. This problem was circumvented later in *live-wire-on-the-fly* (LWOF), by exploiting key properties of Dijkstra's algorithm to determine optimum paths [9]. Several approaches further extended live wire to cope with multiple challenges [10,11].

Methods based on region constraints usually have the advantage of being more easily extended to 3D images. Some popular approaches based on internal and external markers are watershed [12,13] (WS), fuzzy connectedness [14,15] (FC), and the traditional graph cuts [3,16] (GC). These methods can define the object as some optimum cut in the graph and can produce similar results under certain conditions [17,13,15]. However, they may differ in computational efficiency and user involvement, time, and control, depending on the quality of the arc-weight assignment and algorithm chosen for implementation. WS and FC are more robust to the markers'position and better perform in the case of complex object silhouettes (with protrusions and indentations) than GC and LWOF [17,15] (Figure 1a). However, in the presence of poorly defined parts of the boundary (bad arc-weight assignment), they present a leaking problem where parts of the background (object) are conquered by object (background) markers. On the other hand, GC and LWOF produce smoother borders and better perform on the poorly defined parts of the boundary (Figure 1b). This makes interesting to investigate hybrid approaches that can combine the complementary strengths of both paradigms and eliminate their weaknesses.

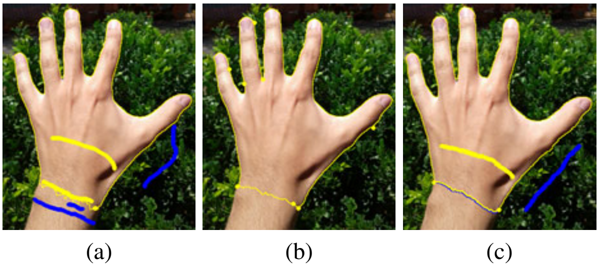


Fig. 1. (a) DIFT-SC handles complex object shapes but suffers from the same drawbacks of FC and WS towards weak boundary information, thus requiring a few markers around the wrist to finish segmentation. (b) Hand segmentation using LWOF with eight anchor points. (c) The hand segmentation by Live Markers used a couple of internal and external markers and one LWOF segment on the wrist to segment the hand.

In this work, we propose a hybrid approach using the above strategy and a common graph-based framework to develop methods, named *Image Foresting Transform* (IFT) [18]. In this framework, an image is interpreted as a graph whose image elements (pixels, vertices, edges, regions) are the nodes and the arcs are defined by some *adjacency relation* between them. A *connectivity function* is assigned to any path in the graph, including trivial paths formed by single nodes. Considering an initial connectivity map with only trivial paths, its maxima (minima) are taken as root nodes. These roots may offer better paths to their adjacent nodes and the adjacent nodes may also propagate better paths in such a way that an optimum-path propagation process transforms the image into an *optimum-path forest*. Different image operators are then reduced to a local processing on the attributes of the forest (optimum paths, root labels, optimum connectivity values). In image segmentation, the IFT has been used to better understand the differences among WS algorithms [12], to considerably speed up FC computation [15], and to create new methods [19].

The IFT algorithm is an extension of Dijkstra's algorithm for multiple sources and more general connectivity functions, whose linear-time implementation is possible in most cases. By adding/removing trees, the optimum-path forest can also be updated in sublinear time with the differential IFT (DIFT) algorithm. This strongly favors the use of the DIFT for multidimensional interactive segmentation with region constraints, such as the method DIFT with seed competition (DIFT-SC) where the seeds come from the internal and external markers [4]. Besides, the time complexity of the algorithm does not increase with the number of objects. Seeds compete among themselves and each object is defined by pixels more strongly connected to its internal seeds than to any other.

Therefore, DIFT-SC and LWOF are combined into a new method called *Live Markers* (LM) in order to reduce user involvement and time for interactive segmentation. LM combines their strengths and eliminate their weaknesses at the same time, by transforming optimum boundary segments from LWOF into internal and external markers for DIFT-SC (Figure 1c). It allows linear-time execution in the first interaction and sublinear-time corrections in the subsequent ones. Although DIFT-SC can use different connectivity functions and handle multiple objects, we will present it for binary segmentation using a connectivity function suitable for complex object silhouettes. As a consequence of that, LM will present the leaking problem on weaker parts of the boundary which will be solved by LWOF or by additional internal and external markers (Figure 1c). We will also present LM for the 2D case. However, it should be clear that its extension to the 3D case is straightforward. LWOF can execute in a few selected slices but object delineation will always be done by DIFT-SC in 3D.

Given that IFT is used in every aspect of LM, some general concepts about its operators and image-derived graphs are presented in Section 2. Section 3 details the segmentation process by Live Markers, including how the object is extracted by DIFT-SC using both manual and LWOF border markers. Lastly, in Section 4 Live Markers is compared with DIFT-SC and LWOF and our conclusions are stated.

2 Image Foresting Transform

An image $\hat{I}$ is a pair $(D_{\hat{I}}, \boldsymbol{I})$, where $D_{\hat{I}} \subset Z^n$ corresponds to the image domain and $\boldsymbol{I}(t)$ assigns a set of m scalars $I_b(t)$, $b = 1, 2, \ldots, m$, to each pixel $t \in D_{\hat{I}}$. The subindex b is removed when $m = 1$. A graph $(\mathcal{N}, \mathcal{A})$ may be defined by taking a set $\mathcal{N} \subseteq D_{\hat{I}}$ of pixels as nodes and an *adjacency relation* $\mathcal{A}$ between nodes of $\mathcal{N}$ to form the arcs. We use $t \in \mathcal{A}(s)$ or $(s, t) \in \mathcal{A}$ to indicate that a node $t \in \mathcal{N}$ is adjacent to a node $s \in \mathcal{N}$. Live Markers considers for LWOF and DIFT-SC a graph in which every pixel is a node and the arcs are defined between the 8 neighbors of a pixel in the image domain (i.e., $\mathcal{N} = D_{\hat{I}}$ and $(s,t) \in \mathcal{A}$ if $\|t - s\| \leq \sqrt{2}$). The arcs $(s,t) \in \mathcal{A}$ are assigned fixed weights $0 \leq w(s,t) \leq K$ computed from local image features and object information [20], such that higher arc weights are assigned across the object's boundary for DIFT-SC (and lower arc weights $\bar{w}(s,t) = K - w(s,t)$ on the object's border for LWOF).

A *path* $\pi_t = \langle t_1, t_2, \ldots, t\rangle$ is a sequence of adjacent nodes with terminus at a node t. A path $\pi_t = \pi_s \cdot \langle s,t\rangle$ indicates the extension of a path π_s by an arc (s,t) and a path $\pi_t = \langle t\rangle$ is said *trivial*. A connectivity function f assigns to any path π_t a value $f(\pi_t)$. A path π_t is optimum if $f(\pi_t) \leq f(\tau_t)$ for any other path τ_t in $(\mathcal{N}, \mathcal{A})$. Considering all possible paths with terminus at each node $t \in \mathcal{N}$, an optimum connectivity map $V(t)$ is created by $V(t) = \min_{\forall \pi_t \, in \, (\mathcal{N},\mathcal{A})}\{f(\pi_t)\}$.

The IFT solves the minimization problem above by computing an *optimum-path forest* — a function P which contains no cycles and assigns to each node $t \in \mathcal{N}$ either its predecessor node $P(t) \in \mathcal{N}$ in the optimum path or a distinctive marker $P(t) = nil \notin \mathcal{N}$, when $\langle t\rangle$ is optimum (i.e., t is said *root* of the forest). The root $R(t)$ of each pixel t can be obtained by following its optimum path backwards in P. However, it is more efficient to propagate them on-the-fly, creating a root map R. Also, the path optimality is only guaranteed for *smooth* functions [18], such as the ones below used for region and boundary based segmentation, respectively

$$\begin{aligned} f_{\max}(\langle t\rangle) &= H(t) \\ f_{\max}(\pi_s \cdot \langle s,t\rangle) &= \max\{f_{\max}(\pi_s), w(s,t)\}, \end{aligned} \tag{1}$$

$$\begin{aligned} f_{\Sigma}^{\circlearrowleft}(\langle t\rangle) &= H(t) \\ f_{\Sigma}^{\circlearrowleft}(\pi_s \cdot \langle s,t\rangle) &= \begin{cases} f_{\Sigma}^{\circlearrowleft}(\pi_s) + \bar{w}^{\beta}(s,t) & \text{if } O(l) \geq O(r) \\ f_{\Sigma}^{\circlearrowleft}(\pi_s) + K^{\beta} & \text{otherwise,} \end{cases} \end{aligned} \tag{2}$$

where $H(t)$ is a *handicap* value specific to each IFT-based operator, l and r are the pixels at the left and right sides of arc $\langle s,t\rangle$, O is a reference map expected to be brighter inside the object, and the parameter $\beta \geq 1$ produces *longer segments* for LWOF in an anti-clockwise fashion. Function $f_{\max}$ propagates the maximum arc weight value along the path, while $f_{\Sigma}^{\circlearrowleft}$ is the oriented additive path-cost function. Other connectivity functions are discussed in [18] for several image operators.

3 Live Markers

In Live Markers, the object is always extracted using optimum seed competition through the DIFT-SC operator [4]. DIFT-SC computes an optimum-path forest spanning from a set of selected marker pixels (seeds) to every node in a graph derived from the image (Section 2). The object is defined as the union of trees rooted at the internal seeds. These seeds can be strokes drawn by the user and/or automatically generated sets of pixels surrounding border segments computed by LWOF. User-drawn markers are often placed at locations with homogeneous color and texture, and may provide useful information for arc-weight estimation. Intelligent arc-weight estimation [20] simplifies the user interface by determining when and where the weights can be recomputed, using an automatically selected subset of the marked pixels with the most representative image attributes that help distinguish object from background. It aims at computing higher arc weights across the object's border than anywhere else, such that the object can be extracted using DIFT-SC with $f_{\max}$ from only two marker pixels, one inside and one outside it [6]. Nevertheless, perfect arc-weight assignment is often not possible and more markers should be placed around weaker parts of the boundary for correction. In this case, the automatic generation of markers surrounding LWOF border segments forms perfect barriers that are much more effective. In fact, LWOF border markers are so important that in many cases they can virtually replace user-drawn markers altogether, since DIFT-SC labels the rest of the image accordingly (e.g., fish in Figure 2).

3.1 Segmentation by Live-Wire-on-the-Fly

Boundary-based segmentation by live wire [1] outputs a closed contour computed as an optimum curve that is constrained to pass through a sequence $\langle s_1, s_2, \ldots, s_N \rangle$ of N anchor points (seeds) selected by the user on the object's boundary, in that order, starting from s_1 and ending in s_N, where $s_1 = s_N$. The optimum curve that satisfies those constraints consists of $N - 1$ segments $\pi_{s_2}, \pi_{s_3}, \ldots, \pi_{s_N}$, where each π_{s_i} is an optimum path connecting s_{i-1} to s_i. Therefore, we can solve this problem by $N - 1$ executions of the IFT and the optimum contour can be obtained from the predecessor map P after the last execution.

To select a new anchor point s_i, the user moves the mouse's cursor and the optimum-path from s_{i-1} to the cursor's position (candidate for s_i) is displayed in real-time. In this work, each execution $i = 2, 3, \ldots, N$ of IFT for live wire uses the initial point s_{i-1} as seed and the oriented version of the additive path-cost function in Eq. 2 (with $H(t) = 0$ if $t = s_{i-1}$, and $H(t) = +\infty$ otherwise). For our purpose, O is taken as the *object membership map* M computed during arc weight estimation [20] as a result of supervised fuzzy pixel classification.

At each IFT iteration ($i = 2, 3, \ldots, N$), the previous segments $\pi_{s_2}, \pi_{s_3}, \ldots, \pi_{s_N}$ are kept unchanged during the algorithm, so their nodes can not be revisited or reseted. Live-wire-on-the-fly [9] is finally obtained by exploiting the Bellman's principle for early termination and incremental computation in each execution of IFT.

3.2 Combination of Live-Wire-on-the-Fly with DIFT-SC

Instead of defining a closed contour to delineate the object, the combination between LWOF and DIFT-SC transforms each optimum boundary segment π_{s_i}, computed by LWOF between two anchor points s_{i-1} and s_i, into region constraints for DIFT-SC. The addition of a new border segment causes DIFT-SC to be instantaneously issued to update the result on-the-fly. Segmentation may continue by prolonging the current border segment, by restarting LWOF at another location with a new anchor point, or by adding/removing markers.

Let $\mathcal{M}$ be a set of marker pixels drawn by the user and $\mathcal{B}$ be the set of pixels that belong to the optimum path π_{s_i} rooted at the anchor point s_{i-1}. The seed set used to extract the object by DIFT-SC can be taken as $\mathcal{U} = \mathcal{M} \cup \mathcal{B} \cup \mathcal{E} \cup \mathcal{D}$, where $\mathcal{E}$ and $\mathcal{D}$

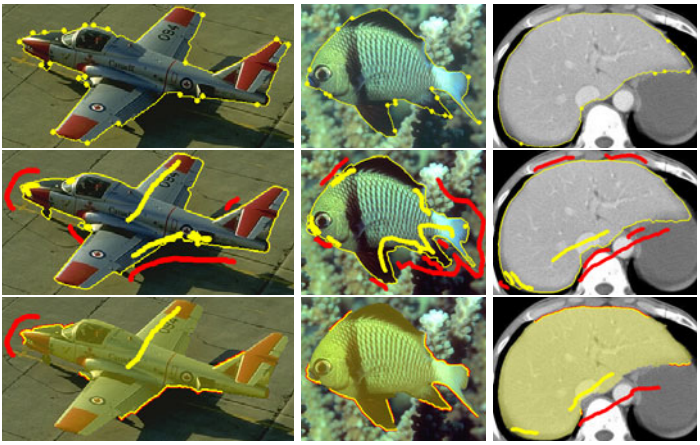


Fig. 2. The images in the first row present the segmentation result using LWOF. The second and third row depict, respectively, the delineation of objects by DIFT and Live Markers. Note how the fish segmentation only required border markers when using LM.

Table 1. Measures of accuracy (F-measure and $\overline{ED}$) and interaction (number of markers for LM and DIFT-SC, and anchor points for LWOF) from both experiments

	GrabCut Dataset			Liver CT Dataset		
	LM	**LWOF**	**DIFT-SC**	**LM**	**LWOF**	**DIFT-SC**
F-measure	98.9 ± 0.5	98.7 ± 0.6	99.0 ± 0.4	98.6 ± 0.2	98.6 ± 0.2	98.5 ± 0.5
$\overline{\textbf{ED}}$	0.6 ± 0.2	0.7 ± 0.2	0.6 ± 0.3	1.3 ± 0.2	1.2 ± 0.1	1.3 ± 0.3
Interactions	6.4 ± 4.1	13.5 ± 8.7	9.6 ± 6.0	10.3 ± 2.3	17.9 ± 5.5	13.9 ± 3.0

are the 8 adjacent pixels to the left and right of π_{s_i}, respectively. The marker label for pixels in $\mathcal{E}$ and $\mathcal{D}$ can be easily determined according to the current orientation being used for LWOF. For instance, in anti-clockwise orientation every $t \in \mathcal{E}$ is assigned an object label $\lambda(t) = 1$, while every $s \in \mathcal{D}$ is given a background label $\lambda(s) = 0$. All nodes in $\mathcal{B}$ are always object markers. Such definition of LWOF border markers ensures a tight seed assignment that protects weaker parts of the boundary.

3.3 Object Extraction by DIFT-SC

Object extraction is performed on an 8-neighbor graph derived from the image $(D_{\hat{I}}, \mathcal{A})$, with regular arc weights $w(s,t)$. All pixels in set $\mathcal{U}$ are taken as seeds for optimum competition by IFT. It is expected that the optimum-path forest P computed on $(D_{\hat{I}}, \mathcal{A})$ for $f_{\max}$ in Eq. 1 with $H(t) = 0$ if $t \in \mathcal{U}$, and $H(t) = +\infty$, otherwise, extracts the object as the union of trees rooted at the object pixels in $\mathcal{U}$. The object is identified as the pixels 1 after a local operation, which assigns the correct label $L(t) = \lambda(R(t)) \in \{0,1\}$ of the root to each pixel $t \in D_{\hat{I}}$. That is, object and background seeds will compete with each other to conquer their most strongly connected pixels, hopefully from the same label.

The user may draw new markers, add LWOF border markers, and/or remove markers by clicking on them, and the optimum-path forest P can be recomputed in a differential way, taking time proportional to the number of pixels in the modified image regions (sublinear time in practice, if $w(s,t)$ is normalized within an integer range of numbers). This approach comprises the differential image foresting transform with seed competition (DIFT-SC) [4]. That is, each marker pixel added to $\mathcal{U}$ may define a new optimum-path tree by invading the trees of other roots. The removal of a marker eliminates all optimum-path trees rooted at it, making their pixels available for a new dispute among the remaining roots.

4 Experiments and Conclusions

Some examples of segmentation using Live Markers are presented in Figure 2. Live Markers was used to segment 22 natural images selected from the GrabCut [16] dataset, and a set with 20 CT-images of the liver from 10 different subjects. These images were also segmented using DIFT-SC and LWOF separately, on graphs whose arc weights were also estimated using the intelligent approach in [20].

Segmentation accuracy is established taking into account region and boundary based metrics (Table 1). Namely, the *F-measure* score computed over the groundtruths and the average euclidean distance between the segmentation masks and groundtruth boundaries $\overline{ED}$. The amount of user interaction is measured by the total number of markers used for LM and DIFT-SC, and the number of anchor points for LWOF (Table 1). While the overall user time spent in segmentation for all methods ranges from 1 to 2 minutes, being greater for LWOF in general, their computational time is low, in the order of 0.2 to 0.6 seconds per interaction for images sized between 481×321 and 640×480 pixels. All experiments were executed in a machine with a 2.2 GHz Intel Core i5 processor and 4 GB of RAM. From our experiments we can see that Live Markers achieves high accuracy, and demands from 26% to 52% less user interaction than LWOF and DIFT-SC used separately. The metrics based on boundary distance indicate some discrepancies between the segmentation mask and the groundtruths. This is mostly related to some discretization artifacts and the manual generation of groundtruths, which produced rough borders. Nevertheless, Live Markers is a promising approach that can be straightforwardly extended to 3d to further reduce user interaction in segmentation for medical image analysis. Future work also involves the development of matting algorithms to produce smoother borders and cope with fine features such as hair.

We would like to thank FAPESP (2009/11908-8, 2011/01434-9, 2009/16428-4, 2007/52015-0), CNPq (481556/2009-5, 201732/2007-6, 302617/2007-8), and the University of Campinas for the financial support, J. K. Udupa and the creators of the GrabCut and Berkeley datasets for the images and groundtruths.

References

1. Falcão, A.X., Udupa, J.K., Samarasekera, S., Sharma, S., Hirsch, B.E., Lotufo, R.A.: User-steered image segmentation paradigms: Live-wire and live-lane. Graphical Models and Image Processing 60, 233–260 (1998)

2. Mortensen, E., Barrett, W.: Interactive segmentation with intelligent scissors. Graphical Models and Image Processing 60, 349–384 (1998)
3. Boykov, Y., Funka-Lea, G.: Graph cuts and efficient N-D image segmentation. Intl. Journal of Computer Vision 70, 109–131 (2006)
4. Falcão, A.X., Bergo, F.P.G.: Interactive volume segmentation with differential image foresting transforms. IEEE Trans. on Medical Imaging 23, 1100–1108 (2004)
5. Bai, X., Sapiro, G.: Geodesic matting: A framework for fast interactive image and video segmentation and matting. Intl. Journal of Computer Vision 82, 113–132 (2009)
6. Miranda, P.A.V., Falcão, A.X., Udupa, J.K.: Synergistic arc-weight estimation for interactive image segmentation using graphs. Computer Vision and Image Understanding 114, 85–99 (2010), doi:10.1016/j.cviu.2009.08.001.
7. Yang, W., Cai, J., Zheng, J., Luo, J.: User-friendly interactive image segmentation through unified combinatorial user inputs. IEEE Trans. on Image Processing 19, 2470–2479 (2010)
8. Martelli, A.: Edge detection using heuristic search methods. Computer Graphics and Image Processing 1, 169–182 (1972)
9. Falcão, A.X., Udupa, J.K., Miyazawa, F.K.: An ultra-fast user-steered image segmentation paradigm: Live-wire-on-the-fly. IEEE Trans. on Medical Imaging 19, 55–62 (2000)
10. Malmberg, F., Vidholm, E., Nystrom, I.: A 3D live-wire segmentation method for volume images using haptic interaction. In: Kuba, A., Nyúl, L.G., Palágyi, K. (eds.) DGCI 2006. LNCS, vol. 4245, pp. 663–673. Springer, Heidelberg (2006)
11. Liu, J., Udupa, J.: Oriented active shape models. IEEE Trans. on Medical Imaging 28, 571–584 (2009)
12. Audigier, R., Lotufo, R.: Watershed by image foresting transform, tie-zone, and theoretical relationship with other watershed definitions. In: Proceedings of the 8th Intl. Symposium on Mathematical Morphology and its Applications to Signal and Image Processing (ISMM), Rio de Janeiro, Brazil, MCT/INPE, pp. 277–288 (2007)
13. Cousty, J., Bertrand, G., Najman, L., Couprie, M.: Watershed cuts: Thinnings, shortest path forests, and topological watersheds. IEEE Trans. on Pattern Analysis and Machine Intelligence 32, 925–939 (2010)
14. Udupa, J., Saha, P., Lotufo, R.: Relative fuzzy connectedness and object definition: Theory, algorithms, and applications in image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence 24, 1485–1500 (2002)
15. Ciesielski, K.C., Udupa, J.K., Falcão, A.X., Miranda, P.A.V.: Fuzzy connectedness and graph cut image segmentation: similarities and differences. In: Proceedings of SPIE on Medical Imaging: Image Processing (to appear, 2011)
16. Rother, C., Kolmogorov, V., Blake, A.: "grabcut": Interactive foreground extraction using iterated graph cuts. ACM Trans. on Graphics 23, 309–314 (2004)
17. Miranda, P.A.V., Falcão, A.X.: Links between image segmentation based on optimum-path forest and minimum cut in graph. Journal of Mathematical Imaging and Vision 35, 128–142 (2009), doi:10.1007/s10851-009-0159-9.
18. Falcão, A.X., Stolfi, J., Lotufo, R.A.: The image foresting transform: Theory, algorithms, and applications. IEEE Trans. on Pattern Analysis and Machine Intelligence 26, 19–29 (2004)
19. Miranda, P.A.V., Falcão, A.X., Spina, T.V.: The Riverbed approach for user-steered image segmentation. In: 18th International Conference on Image Processing (ICIP), Brussels, Belgium (to appear, 2011)
20. Spina, T.V., Falcão, A.X.: Intelligent understanding of user input applied to arc-weight estimation for graph-based foreground segmentation. In: Proceedings of the XXIII Conference on Graphics, Patterns and Images (SIBGRAPI), Gramado, Brazil. IEEE, Los Alamitos (2010)

Kernelising the Ihara Zeta Function

Furqan Aziz, Richard C. Wilson, and Edwin R. Hancock*

Department of Computer Science, The University of York, YO10 5GH, UK
{furqan,wilson,erh}@cs.york.ac.uk

Abstract. The Ihara Zeta Function, related to the number of prime cycles in a graph, is a powerful tool for graph clustering and characterization. In this paper we explore how to use the Ihara Zeta Function to define graph kernels. We propose to use the coefficients of reciprocal of Ihara Zeta Function for defining a kernel. The proposed kernel is then applied to graph clustering.

Keywords: Ihara Zeta Function, Graph Kernels, Cycle Kernels.

1 Introduction

Although originally developed for vector-data, there has recently been a concerted effort to extend kernel methods to the structural domain, i.e. to measuring the similarity of strings, trees and graphs. While there is an order relation in strings and trees, graphs represent more difficult structures to kernelise since there is no order relation. It is for this reason that the construction of graph kernels has proved to be particularly challenging.

The lack of an order relation renders the problem both computationally and algorithmically burdensome. For instance subgraph isomorphism is known to be NP-complete, and this makes the exact solution computationally intractable. It is for this reason that inexact and decomposition methods have been used instead. Methods falling into the former category include the use of approximate methods to compute graph-edit distance[1,16], and those falling into the latter category include those that decompose a graph into path length[3], random walk[2] and cycle kernels [5,4]. One of the advantages of the decomposition methods is that they lead to kernels that can be computed in polynomial time.

One of the most popular polynomial time graph kernels is the random walk kernel[2]. This relies on the fact that the path length distribution of a graph can be easily computed from powers of the adjacency matrix, and this is an operation that can be computed in polynomial time. Moreover, by using a product graph formalism the computation can be accelerated[2]. There are a number of well documented problems with the random walk kernel. These include a)problems that different graphs are mapped to the same point in the feature-space of the random walk (this can be attributed to the cospectrality of graphs), and b)the

* Edwin Hancock is supported by the EU FET project SIMBAD and by a Royal Society Wolfson Research Merit Award.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 219–227, 2011.

fact that random walks totter and may visit the same edges and nodes multiple times. Both of these problems mean that the ability of the graph kernel to discriminate graphs of different structure is reduced. One way to overcome these problems is to define kernels on cycles rather than paths (walks)[5].

Recently Peng et al[6,7,8] have explored the use of the Ihara zeta function as a mean of gauging cycle structure in graphs. The Ihara zeta function is computed by first converting a graph into the equivalent oriented line digraph, and then computing the characteristic polynomial of the resulting structure. The coefficients of the characteristic polynomials are related to the frequencies of prime cycles of different size, and can be computed in polynomial time from the eigenvalues of the oriented line-digraph adjacency matrix. The method can be easily extended from simple graphs to both weighted graphs and hypergraphs. Moreover, since the method is based on a oriented line digraph, it is closely akin to the discrete time quantum walk on a graph[14] and is hence less prone to problems of failing to distinguish graphs due to cospectrality of the Laplacian or adjacency matrices.

The aim in this paper is to explore how to use the Ihara zeta function to define a new family of graph kernels based on cycle composition. There are a number of ways in which this can be achieved, however here we choose to define our kernel using the coefficients of reciprocal of the Ihara zeta function. We show how to compute these coefficients efficiently using the complete Bell polynomials. The resulting kernel is compared to the path length kernel, and evaluated on graphs extracted from image data.

2 The Ihara Zeta Function

The Ihara zeta function associated to a finite connected graph G is defined to be a function of $u \in \mathbb{C}$ with u sufficiently small by [15]

$$\zeta_G(u) = \prod_{c \in [C]} \left(1 - u^{l(c)}\right)^{-1} \qquad (1)$$

The product is over equivalence classes of primitive closed backtrackless, tailless cycles $c = (v_1, v_2, v_3, ..., v_m = v_1)$ of positive length m in G. Here $l(c)$ *length of* $c =$ number of edges in c.

The Ihara zeta Function can also be written in the form of a determinant expression [10]

$$\zeta_G(u) = \frac{1}{\det(I - uT)} \qquad (2)$$

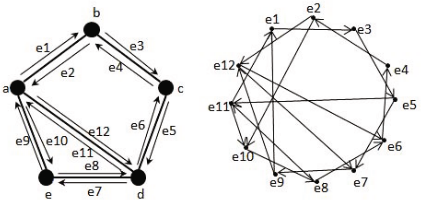


Fig. 1. Example of an undirected graph with five nodes and six edges and its oriented line digraph

where T, the Perron-Frobenius operator, is the adjacency matrix of the oriented line graph of the original graph. The size of T is $2m \times 2m$, where m is the number of edges in original graph. I is the identity matrix of size $2m$. The oriented line graph is constructed by taking two nodes corresponding to each edge in the graph. A node corresponding to edge i is connected to a node corresponding to edge j, if edge i feeds into edge j to form a no backtrack path. Figure 1 shows a graph and its oriented line graph.

3 The Coefficients of Reciprocal of Ihara Zeta Function

The reciprocal of the Ihara zeta function can be written in terms of a determinant of the matrix T, and hence in the form of a polynomial of degree 2m:

$$\zeta_G(u)^{-1} = \det(I - uT) = c_0 + c_1 u + c_2 u^2 + c_3 u^3 + ... + c_{2m} u^{2m} \tag{3}$$

These coefficients are related to the number of prime cycles in the graph. If G is a simple graph, the coefficients c_3, c_4 and c_5 are the negative of twice the number of triangles, squares, and pentagons in G respectively. The coefficient c_6 is the negative of the twice the number of hexagons in G plus four times the number of pairs of edge disjoint triangles plus twice the number of pairs of triangles with a common edge, while c_7 is the negative of the twice the number of heptagons in G plus four times the number of edge disjoint pairs of one triangle and one square plus twice the number of pairs of one triangle and one square that share a common edge[9]. The highest order coefficient is associated with the number of edges incident to vertex v_i, i.e., the node degree $d(v_i)$[10]:

$$c_{2m} = (-1)^{|V|G|-E|G||} \prod_{v_i \in V} (d(v_i) - 1) \tag{4}$$

The above coefficients can be computed as a summation of a series of determinants[11,6]

$$c_n = \sum_{\binom{2m}{2m-k}} \det \begin{pmatrix} b_{1,1} & b_{1,2} & ... & b_{1,2m} \\ b_{2,1} & b_{2,2} & ... & b_{2,2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{2m,1} & b_{2m,2} & ... & b_{2m,2m} \end{pmatrix}$$

This method, however, computes $\binom{2m}{2m-k}$ determinants of size $2m \times 2m$ to find one coefficient c_k. Here we derive a new method for computing the coefficients of the reciprocal of Ihara zeta function, which requires only the computation of the determinant of one matrix. The Ihara zeta function can formally be written in terms of a power series of the variable u, by[13]

$$\zeta_G(u) = \exp\left(\sum_{m \geq 1} \frac{\mathrm{Tr}\,(T^m)}{m} u^m\right) \tag{5}$$

where $\mathrm{Tr}\,(T^m)$ is the trace of T^m. Using (3) and (5), we get

$$\sum_{m \geq 0} c_m u^m = \exp\left(-\sum_{m \geq 1} \frac{\mathrm{Tr}\,(T^m)}{m} u^m\right) \tag{6}$$

The coefficient c_k, can be computed by evaluating the k^{th} derivative of Equation(6) at u=0. The first five coefficients are

$c_0 = 1, c_1 = -\mathrm{Tr}[T], c_2 = \frac{1}{2!}\left(-\mathrm{Tr}[T^2] + \mathrm{Tr}[T]^2\right)$,
$c_3 = \frac{1}{3!}\left(-2\mathrm{Tr}[T^3] + 3\mathrm{Tr}[T^2]\mathrm{Tr}[T] - \mathrm{Tr}[T]^3\right)$,
$c_4 = \frac{1}{4!}\left(-6\mathrm{Tr}[T^4] + 8\mathrm{Tr}[T^3]\mathrm{Tr}[T] + 3\mathrm{Tr}[T^2]^2 - 6\mathrm{Tr}[T^2]\mathrm{Tr}[T]^2 + \mathrm{Tr}[T]^4\right)$

In general the n^{th} coefficient is given by

$$c_n = \sum_{k_1,k_2,...,k_n} \left(-\frac{x_1}{1}\right)^{k_1} \left(-\frac{x_2}{2}\right)^{k_2} ... \left(-\frac{x_n}{n}\right)^{k_n} \tag{7}$$

where $k_1 + 2k_2 + 3k_1 + ... + nk_n = n$ and $x_k = -(k-1)!Tr[T^k]$. We can write c_n in terms of Bell polynomials[12]:

$$c_n = \frac{1}{n!}\left(\sum_{k=1}^{n} B_{n,k}\left(x_1, x_2, ..., x_{n-k+1}\right)\right) \tag{8}$$

$$= \frac{1}{n!} B_n\left(x_1, x_2, ..., x_n\right) \tag{9}$$

where $B_{n,k}\left(x_1, x_2, ..., x_{n-k+1}\right)$ are partial Bell polynomials and $B_n\left(x_1, x_2, ..., x_n\right)$ is the complete Bell polynomial. Since the complete Bell polynomial can be written in the form of a determinant, we can, therefore, write c_n as

$$c_n = \frac{1}{n!}\det\begin{pmatrix} x_1 & \binom{n-1}{1}x_2 & \binom{n-1}{2}x_3 & ... & x_n \\ -1 & x_1 & \binom{n-2}{1}x_2 & ... & x_{n-1} \\ 0 & -1 & x_1 & ... & x_{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & ... & x_1 \end{pmatrix}$$

Each x_k can be efficiently computed as

$$x_k = -(k-1)!\left(\sum \lambda_i^k\right) \tag{10}$$

where $\lambda_1, \lambda_2, \lambda_3, ...$ are the distinct eigenvalues of T.

Since $\mathrm{Tr}[T^k]$ is the number of all prime cycles of length k[13], the above expression for the coefficients gives us some interesting information about each coefficient. For example, when G is a simple graph then

- $c_0 = 1$
- $c_1 = -\mathrm{Tr}[T] = 0$. Since there are no loops in a simple graph so $\mathrm{Tr}[T]$ is always zero.
- $c_2 = \frac{1}{2!}\left(-\mathrm{Tr}[T^2] + \mathrm{Tr}[T]^2\right) = 0$. Since there are no cycles of length two in a simple graph, so $\mathrm{Tr}[T^2]$ is always zero.
- $c_3 = \frac{1}{3!}\left(-2\mathrm{Tr}[T^3] + 3\mathrm{Tr}[T^2]\mathrm{Tr}[T] - \mathrm{Tr}[T]^3\right) = -\frac{1}{3}\mathrm{Tr}[T^3]$. So c_3 depends only on the number of triangles in the graph.
- $c_4 = -\frac{1}{4}\mathrm{Tr}[T^4]$. So c_4 depends only on the number of squares in the graph.
- $c_5 = -\frac{1}{5}\mathrm{Tr}[T^5]$. So c_5 depends only on the number of pentagons in the graph.
- $c_6 = -\frac{1}{6}\mathrm{Tr}[T^6] + \frac{1}{18}\mathrm{Tr}[T^3]^2$. So c_6 depends only on the number of prime cycles of length 6 and the number of triangles in the graph.

4 Graph Kernel

Graph kernel is a positive definite kernel on set of graphs $\mathcal{G}$. For such kernel $\kappa : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ it is known that a map $\Phi : \mathcal{G} \to \mathcal{H}$ into a Hilbert space $\mathcal{H}$ exists, such that $\kappa\left(G, G'\right) = \langle \Phi(G), \Phi(G') \rangle$ for all $G, G' \in \mathcal{G}$[2]. Our objective in this paper is to use the graph kernel to measure the similarity between graphs.

Gärtner et al [2] have defined graph kernel using random walk, which is based on the idea of counting the number of matching walks in two input graphs. Their kernel for the two input graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ is given by the direct product graph $G_\times$:

$$\kappa_\times\left(G_1, G_2\right) = \sum_{i,j=1}^{|V_\times|} \left[\sum_{n=0}^{\infty} \epsilon_n A_\times^n \right] \tag{11}$$

where $A_\times$ is the adjacency matrix of $G_\times = (V_\times, E_\times)$, which is defined as
$V_\times(G_1 \times G_2) = \{(v_1, v_2) \in V_1 \times V_2 : label(v_1) = label(v_2)\}$
$E_\times(G_1 \times G_2) = \{\left(\left(u_1, u_2\right), \left(v_1, v_2\right)\right) \in V^2\left(G_1 \times G_2\right) :$
$(u_1, v_1) \in E_1 \wedge (u_2, v_2) \in E_2 \wedge label\left(u_1, v_1\right) = label\left(u_2, v_2\right)\}$

An interesting approach to defining a graph kernel is to use T, the adjacency matrix of the oriented line graph, instead of using the adjacency matrix of the original graph. Since the oriented line graph captures the backtrackless structure of a graph, the proposed kernel can be a very good measure for the similarity of graphs. There are however number of problems with kernels based on random walks. Such kernels can be very expensive to compute because the direct product graph can have $|V_1| \times |V_2|$ nodes. Finding the higher power of such matrix is computationally expensive. These kernels can also lead to the problem of tottering. i.e., visiting the same node or edge multiple times. One way to overcome such problems is to use kernels based on the set of all paths or the set of all cycles. Horváth et al [5] have defined kernel based on set of all cyclic patterns in the graph. However their kernel can only be applied to graphs where the number of cycles is bounded by a constant. The reason is that computing such kernels can take exponential time. A number of kernels therefore have been defined that use only a subset of all paths or all cycles. Borgwardt and Kriegel[3] have defined their kernel based on the shortest paths between every pair of nodes in the graph. Qiangrong et al [4] have defined a kernel which is based on the subset of cycles of undirected graph. To compute the kernel, they first find the spanning tree of undirected graph. Using the spanning tree and each edge $e \in \{E(G) - E_{span}(T)\}$ they find the cycles in the graphs. They have used these cycles to define a graph kernel. So their kernel is based on only $||E(G)| - |E_{span}(T)||$ cycles, which is much smaller than the actual number of cycles in the graph.

Here we propose the use of the coefficients of the reciprocal of Ihara zeta function for computing graph kernels. We propose to use the feature vector $v = [\epsilon_3 c_3 \ \epsilon_4 c_4 \ \epsilon_5 c_5 \ ... \ \epsilon_k c_k]$ for the graph G, where c_i is the i^{th} coefficients of the reciprocal of Ihara zeta function of graph G. Since $c_0 = 1$, $c_1 = 0$ and $c_2 = 0$, we have ignored these coefficients in the feature vector. The reason for assigning different weights to different coefficients is that the coefficients besides c_3, c_4

and c_5 provide some redundant information. We choose the weights in such a way that the higher order coefficients are assigned smaller weights. We propose to choose $\epsilon_i = \epsilon^{i-2}$ for $i \geq 3$ and $0 < \epsilon < 1$. The value of ϵ depends on the dataset we are using. In practice we select a smaller value for dense graph and a larger value for sparse graph. This is because for dense graphs, the higher order coefficients add more noise to the structural representation of the graph. Our graph kernel is then given by

$$\kappa\left(G_1, G_2\right) = \sum_{\alpha_i \in v_1, \beta_i \in v_2} \alpha_i \beta i \tag{12}$$

where v_1 and v_2 are the feature vectors that are computed from the graphs G_1 and G_2 respectively, as discussed above. Since our feature vector is constructed using the coefficients of the reciprocal of Ihara zeta function, which are related to the number of prime cycles in the graph, our kernel gives a measure of similarity between graphs. The proposed graph kernel is positive definite, since it is a dot product of two feature vectors. In terms of time complexity, our proposed kernel can be computed in $O(n^3)$, once the eigenvalues of oriented line graph are known. In the case of sparse unlabeled graphs, which is usually the case when the graphs are extracted from images, our method outperforms most of the alternative methods.

5 Experiment

To evaluate the performance of our graph kernel, we choose a graph set, extracted from images of three objects in the COIL dataset. Here 90 images (30 per object) should be classified into 3 distinct classes. These objects are shown in Fig 2. For this dataset we choose $\epsilon = 0.2$.

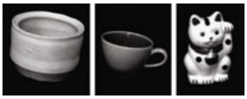


Fig. 2. COIL dataset

We first extract the feature points from the images. For this purpose we use the Harris corner detector[17]. We then construct a Delaunay graph from these feature points as nodes. To visualize the results, we perform PCA on the similarity matrix obtained by applying our kernel to the graphs of the objects. Figure 3 shows the results of our clustering using first three eigenvectors.

To evaluate the performance, we compare our method with both shortest path kernel[3] and random walk kernel[2]. To validate the performance of clustering we use Rand index, which measures the consistency of a given clustering with higher value indicating better clustering. Table 1 shows the rand indices of both

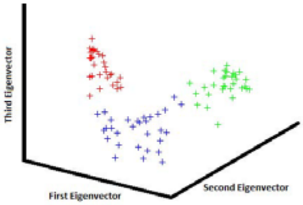


Fig. 3. Performance of Clustering

Table 1. Rand Indices

Method	Rand Index
Shortest path Kernel	0.7336
Random walk kernel	0.8669
Proposed kernel	0.9172

the methods. It is clear from the table that our method is superior to both the shortest path kernel and random walk kernel. We have also compared our method with one that uses a pattern vector from the coefficients of Ihara zeta function [6]. Table 2 shows the rand indices of the two methods using different number of coefficients. These results are also shown in Fig 4. For large feature vector, we normalize the vector to the unit vector. It is clear from the figure that our method gives good results. When using two or three components, the results are comparable. However if we increase the number of components, the result of our method are much better. We have also compared both methods using the coefficients $c_3, c_4, c_5, c_6, c_7, ln\left(|c_{2m}|\right)$, suggested by Peng et al [6]. The results are plotted using '*' in Fig 4, which shows that our method gives good results. It is clear from table 1 that our method using 4 to 6 components gives best results.

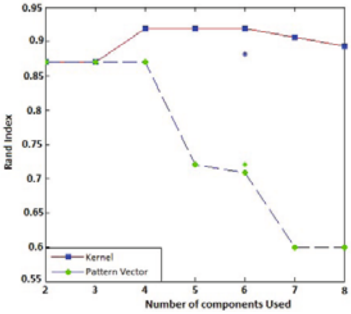


Fig. 4. Number of coefficients Used

Table 2. Rand Indices (Comparison with Pattern Vector)

	2	3	4	5	6	6*	7	8
Pattern Vector	0.8699	0.8699	0.8699	0.7206	0.7081	0.7206	0.6002	0.6002
Kernel	0.8699	0.8699	0.9191	0.9191	0.9191	0.8819	0.9059	0.8913

6 Conclusion

In this paper we have defined a positive definite graph kernel using the coefficients of reciprocal of Ihara Zeta Function. The kernel can be computed very efficiently in polynomial amount of time. We have also derive a new method for computing these coefficients. The proposed scheme is superior to path length kernel, both in terms of time and accuracy, and is better than the one that uses pattern vectors from coefficients of reciprocal of Ihara Zeta function.

References

1. Bunke, H.: On a relation between graph edit distance and maximum common subgraph. Pattern Recognition Lett. 18(8), 689–694 (1997)
2. Gärtner, T., Flach, P., Wrobel, S.: On graph kernels: Hardness results and efficient alternatives. In: Schölkopf, B., Warmuth, M.K. (eds.) COLT/Kernel 2003. LNCS (LNAI), vol. 2777, pp. 129–143. Springer, Heidelberg (2003)
3. Borgwardt, M., Kriegel, H.: Shortest-path kernels on graphs. In: Proceedings of 5th IEEE Internationl Conference on Data Mining (ICDM 2005), pp. 74–81 (2005)
4. Qiangrong, J., Hualan, L., Yuan, G.: Cycle kernel based on spanning tree. In: Proc. of International Conference on Electrical and Control Engineering 2010, pp. 656–659 (2010)
5. Horváth, T., Gärtner, T., Wrobel, S.: Cyclic pattern kernels for predictive graph mining. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery, pp. 158–167 (2004)
6. Ren, P., Wilson, R.C., Hancock, E.R.: Pattern vectors from the Ihara zeta function. In: 19th International Conference on Pattern Recognition, pp. 1–4 (2008)
7. Ren, P., Wilson, R.C., Hancock, E.R.: Graph Characteristics from the Ihara Zeta Function. In: da Vitoria Lobo, N., Kasparis, T., Roli, F., Kwok, J.T., Georgiopoulos, M., Anagnostopoulos, G.C., Loog, M. (eds.) S+SSPR 2008. LNCS, vol. 5342, pp. 257–266. Springer, Heidelberg (2008)
8. Ren, P., Aleksić, T., Wilson, R.C., Hancock, E.R.: Hypergraphs, Characteristic Polynomials and the Ihara Zeta Function. In: Jiang, X., Petkov, N. (eds.) CAIP 2009. LNCS, vol. 5702, pp. 369–376. Springer, Heidelberg (2009)
9. Scott, G., Storm, C.K.: The coefficients of the Ihara Zeta Function. Involve - a Journal of Mathematics 1(2), 217–233 (2008)
10. Kotani, M., Sunada, T.: Zeta function of finite graphs. Journal of Mathematics, University of Tokyo 7(1), 7–25 (200)
11. Brooks, B.P.: The coefficients of the characteristic polynomial in terms of the eigenvalues and the elements of an n×n matrix. Applied Mathematics, 511–515 (2006)

12. Mihoubi, M.: Some congruences for the partial Bell polynomials Journal of Integer Sequences (2009)
13. Mizuno, H., Sato, I.: Zeta functions of diagraphs. Linear Algebra and its Applications 336, 181–190 (2001)
14. Ren, P., Aleksic, T., Emms, D., Wilson, R.C., Hancock, E.R.: Quantum walks, Ihara zeta functions and cospectrality in regular graphs. Quantum Information Processing (in press)
15. Bass, H.: The IharaSelberg zeta function of a tree lattice. Internat. J. Math, 717–797 (1992)
16. Riesen, K., Bunke, H.: Approximate graph edit distance computation by means of bipartite graph matching. Image Vision Comput., 950–959 (2009)
17. Harris, C., Stephens, M.: A combined corner and edge detector. In: Fourth Alvey Vision Conference, Manchester, UK, pp. 147–151 (1988)

A Hypergraph-Based Approach to Feature Selection

Zhihong Zhang and Edwin R. Hancock[*]

Department of Computer Science, University of York, UK
{zhihong,erh}@cs.york.ac.uk

Abstract. In many data analysis tasks, one is often confronted with the problem of selecting features from very high dimensional data. The feature selection problem is essentially a combinatorial optimization problem which is computationally expensive. To overcome this problem it is frequently assumed that either features independently influence the class variable or do so only involving pairwise feature interaction. To overcome this problem, we draw on recent work on hyper-graph clustering to extract maximally coherent feature groups from a set of objects using high-order (rather than pairwise) similarities. We propose a three step algorithm that, namely, i) first constructs a graph in which each node corresponds to each feature, and each edge has a weight corresponding to the interaction information among features connected by that edge, ii) perform hypergraph clustering to select a highly coherent set of features, iii) further selects features based on a new measure called the multidimensional interaction information (MII). The advantage of MII is that it incorporates third or higher order feature interactions. This is realized using hypergraph clustering, which separates features into clusters prior to selection, thereby allowing us to limit the search space for higher order interactions. Experimental results demonstrate the effectiveness of our feature selection method on a number of standard data-sets.

Keywords: Hypergraph clustering, Multidimensional interaction information(MII).

1 Introduction

High-dimensional data pose a significant challenge for pattern recognition. The most popular methods for reducing dimensionality are variance based subspace methods such as PCA. However, the extracted PCA feature vectors only capture sets of features with a significant combined variance, and this renders them relatively ineffective for classification tasks. Hence, it is crucial to identify a smaller subset of features that are informative for classification and clustering. The idea underpinning feature selection is to a) reduce the dimensionality of the feature space, b) speed up and reduce the cost of a learning algorithm, c) obtain

[*] Edwin Hancock is supported by the EU FET project SIMBAD and by a Royal Society Wolfson Research Merit Award.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 228–235, 2011.

the feature subset which is most relevant to classification. In practice, however, optimal feature selection requires 2^n feature subset evaluations, where n is the original number of features and many problems related to feature selection are shown to be NP-hard [2]. Traditional feature selection methods address this issue by partitioning the original feature set into distinct clusters formed by similar features [3]. However, all of the above methods are weakened by only considering pairwise relations. In some applications higher-order relations are more appropriate to the classification task on hand, and approximating them in terms of pairwise interactions can lead to a substantial loss of information.

To overcome the above problem, in this paper, we propose a hypergraph-based approach to feature selection. Hypergraph clustering is capable of detecting high-order feature similarities. In this feature selection scheme, the original features are clustered into different groups based on hypergraph clustering and each group includes just a small set of features. In addition, for each group, a new feature selection criterion referred to as multidimensional interaction information (MII) $I(F;C)$ is applied to feature selection. In contrast to existing feature selection criterion, MII is sensitive to the relations between feature combinations and can be used to seek third or ever higher order dependencies between the relevant features. However, the limitations of the MII criterion are that it requires an exhaustive "combinatorial" search over the feature space and demands estimation of the joint probability distribution for features using large training samples. So most existing works use MII based on the second-order feature dependence assumption [1]. Since hypergraph clustering separates features into clusters in advance, this allows us to limit the search space for higher order interactions directly using the MII criterion $I(F;C)$ for feature selection. Using the Parzen window for probability distribution estimation, we apply a greedy strategy to incrementally select the features that maximize the multidimensional mutual information between the current selected features and the output class set.

2 Hypergraph Clustering Algorithm

Concept of hypergraph: A hypergraph is defined as a triplet $H = (V, E, s)$, where $V = \{1, \dots, n\}$ is the node-set, E is a set of non-empty subsets of V or hyperedges and s is a weight function which associates a real value with each edge. A hypergraph is a generalization of a graph. Unlike graph edges which consisting pairs of vertices, hyperedges are arbitrarily sized sets of vertices. Examples of a hypergraph are shown in Fig. 1. For the hypergraph, the vertex set is $V = \{v_1, v_2, v_3, v_4, v_5\}$, where each vertex represents a feature, and the hyperedge set is $E = \{e_1 = \{v_1, v_3\}, e_2 = \{v_1, v_2\}, e_3 = \{v_2, v_4, v_5\}, e_4 = \{v_3, v_4, v_5\}\}$. The number of vertices constituting each hyperedge represent the order of the relationship between features.

Hypergraph Clustering Algorithm: Let $H = (V, E, s)$ be a hypergraph clustering problem. We can locate the hypergraph cluster by finding the solutions of the following non-linear optimization problem that maximizes the functional

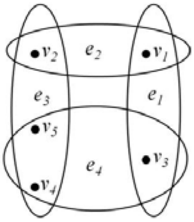


Fig. 1. Hypergraph example

$$f(\mathbf{x}) = \sum_{e \in E} s(e) \prod_{i \in e} x_i \ . \tag{1}$$

subject to $\mathbf{x} \in \triangle$, where $\triangle = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq 0, \sum_{i=1}^{n} x_i = 1\}$ and s is a weight function which associates a real value with each edge. The local maximum of $f(x)$ can be solved using the Baum-Eagon inequality and leads to the iteratively updated lambda:

$$z_i = \frac{x_i \partial_i f(x)}{\sum_{j=1}^{n} x_j \partial_j f(x)}, i = 1, \ldots, n \ . \tag{2}$$

where $f(x)$ is a homogeneous polynomial in the variables x_i and $z = \mathcal{M}(x)$ is a growth transformation of x. The Baum-Eagon inequality $f(\mathcal{M}(x)) > f(x)$ provides an effective iterative means for maximizing polynomial functions in probability domains.

3 Feature Selection Using Hypergraph Clustering

In this paper we aim to utilize the hypergraph clustering algorithm for feature selection. Using a hypergraph representation of the features, there are three steps to the algorithm, namely a) computing the relevance matrix **S** based on the interaction information among feature vectors, b) hypergraph clustering to cluster the feature vectors and c) selecting the optimal feature set from each cluster using the multidimensional interaction information (MII) criterion. In the remainder of this paper we describe these elements of our feature selection algorithm in more detail.

Computing the Relevance Matrix: In accordance with Shannon's information theory, the uncertainty of a random variable Y can be measured by the entropy $H(Y)$. For two variables X and Y, the conditional entropy $H(Y|X)$ measures the remaining uncertainty about Y when X is known. The mutual information (MI) represented by $I(X;Y)$ quantifies the information gain about Y provided by variable X. The relationship between $H(Y)$, $H(Y|X)$ and $I(X;Y)$ is $I(X;Y) = H(Y) - H(Y|X)$. As defined by Shannon, the initial uncertainty for the random variable Y is expressed as: $H(Y) = -\sum_{y \in Y} P(y) \log P(y)$, where

$P(y)$ is the prior probability density function over $y \in Y$. The remaining uncertainty in the variable Y if the variable X is known is defined by the conditional entropy $H(Y|X) = -\int_x p(x)\{\sum_{y\in Y} p(y|x) \log p(y|x)\}dx$, where $p(y|x)$ denotes the posterior probability for variable $y \in Y$ given another random variable $x \in X$. After observing the variable vector x, the amount of additional information gain is given by the mutual information (MI) $I(X;Y) = \sum_{y\in Y} \int_x p(y,x) log \frac{p(y,x)}{p(y)p(x)} dx$.

From the above definition, we can see that mutual information quantifies the information which is shared by two variables X and Y. When the $I(X;Y)$ is large, this implies that variable $x \in X$ and variable $y \in Y$ are closely related, otherwise, when $I(X;Y)$ is equal to 0, this means that two variables are totally unrelated. Analogically, the conditional mutual information of X and Y, denoted as $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$, represents the quantity of information shared by X and Y when Z is known. The conditioning on a third random variable may either increase or decrease the original mutual information.That is, the difference between the conditional mutual information and the simple mutual information, referred to as the Interaction Information is:

$$I(X;Y;Z) = I(X;Y|Z) - I(X;Y) \; . \tag{3}$$

The interaction information measures the influence of the variable Z on the amount of information shared between variables $\{Y, X\}$, the value can be positive, negative, or zero. A zero value means that the relation between X and Y is entirely because of Z. A positive value means that X and Y are independent of each other. However, when combined with Z, X and Y are correlated with each other. A negative value indicates that Z can account for or explain the correlation between X and Y. The extension of interaction information to n variables is defined recursively,

$$I(\{X_1,\ldots,X_n\}) = I(\{X_1,\ldots,X_{n-1}\}|X_n) - I(\{X_1,\ldots,X_{n-1}\}) \; . \tag{4}$$

In our feature selection scheme, the high-order relevance of features is computed using interaction information. Suppose there are N training samples, each having K feature vectors. The k^{th} feature vector for the l^{th} training sample is f_k^l, and so we can represent the k^{th} feature vector for the N training samples as the long vector $F_k = \{f_k^1, f_k^2, \ldots, f_k^N\}$. For three feature vectors F_{k1}, F_{k2} and F_{k3}, their interaction information $I(F_{k1}, F_{k2}, F_{k3})$ can be computed by Equation (3). The relevance degree among three feature vectors F_{k1}, F_{k2} and F_{k3} can be defined as

$$\mathbf{S}(F_{k1}, F_{k2}, F_{k3}) = \frac{3I(F_{k1}, F_{k2}, F_{k3})}{H(F_{k1}) + H(F_{k2}) + H(F_{k3})} \; . \tag{5}$$

where $k1, k2, k3 \in K$ and the higher the value of $\mathbf{S}(F_{k1}, F_{k2}, F_{k3})$ the more relevant are the features F_{k1}, F_{k2} and F_{k3}. Otherwise, if $\mathbf{S}(F_{k1}, F_{k2}, F_{k3}) = 0$, the three features are totally unrelated. In addition, for the above computation, we use Parzen-Rosenblatt window method to estimate the probability density function of random variables F_{k1}, F_{k2} and F_{k3}. The Parzen probability density estimation formula is given by: $p(x) = \frac{1}{N}\phi(\frac{x-x_i}{h})$, where $\phi(\frac{x-x_i}{h})$ is the window function and h is the window width. Here, we use a Gaussian as the window function, so $\phi(\frac{x-x_i}{h}) = \frac{1}{(2\pi)^{\frac{d}{2}} h^d |\Sigma|^{\frac{1}{2}}} \exp(\frac{(x-x_i^T)\Sigma^{-1}(x-x_i)}{-2h^2})$, where Σ is the covariance of $(x - x_i)$, d is the length of vector x. When $d = 1$, $p(x)$ estimates the marginal density and when $d = 3$, $p(x)$ estimates the joint density of variables such as F_{k1}, F_{k2} and F_{k3}.

Hypergraph Clustering: the hypergraph clustering algorithm commences from the relevance matrix and iteratively bi-partitions the features into a foreground cluster and a background cluster. It locates the foreground cluster progressively and hierarchically. The clustering process stops when all the features are grouped into either the foreground or background cluster.

Selecting Key Features: The multidimensional interaction information between feature vector $F = \{f_1, \ldots, f_m\}$ and class variable C is:

$$I(F; C) = \sum_{f_1,\ldots,f_m} \sum_{c \in C} P(f_1, \ldots, f_m; c) \times \log \frac{P(f_1, \ldots, f_m; c)}{P(f_1, \ldots, f_m)P(c)} \ . \qquad (6)$$

The main reason for using $I(F; C)$ as a feature selection criterion is that since $I(F; C)$ is a measure of the reduction of uncertainty in class C due to knowledge of the feature vector $F = \{f_1, \ldots, f_m\}$, from an information theoretic perspective selecting features that maximize $I(F; C)$ translates into selecting those features that contain the maximum information about class C. In practice, and as noted in the introduction, locating a feature subset that maximizes $I(F; C)$ presents two problems: 1) it requires an exhaustive "combinatorial" search over the feature space, and 2) it demands large training sample sizes to estimate the higher order joint probability distribution in $I(F; C)$ with a high dimensional kernel [6]. Bearing these obstacles in mind, most of the existing related papers approximate $I(F; C)$ based on the assumption of lower-order dependencies between features. For example, the first-order class dependence assumption includes only first-order interactions. That is, it assumes that each feature independently influences the class variable, so as to select the *mth* feature, f_m, $P(f_m|f_1, \ldots, f_{m-1}, C) = P(f_m|C)$. A second-order feature dependence assumption is proposed by Guo and Nixon [5] to approximate $I(F; C)$, and this is arguably the most simple yet effective evaluation criterion for selecting features. The approximation is given as

$$I(F; C) \approx \widehat{I}(F; C) = \sum_i I(f_i; C) - \sum_i \sum_{j>i} I(f_i; f_j) + \sum_i \sum_{j>i} I(f_i; f_j|C) \ . \qquad (7)$$

Although an MII based on the second-order feature dependence assumption can select features that maximize class-separability and simultaneously minimize dependencies between feature pairs, there is no reason to assume that the final optimal feature subset is formed by pairwise interactions between features. Infact, it neglects the fact that third or higher order dependencies can be lead to an optimal feature subset.

The primary reason for using the approximation $\widehat{I}(F;C)$ for feature selection instead of directly using multidimensional interaction information $I(F;C)$ is that $I(F;C)$ requires estimation of the joint probability distribution of features using a large training sample. Consider the joint distribution $P(F) = P(f_1, \ldots, f_m)$, by the chain rule of probability

$$P(f_i, \ldots, f_m) = P(f_1)P(f_2|f_1) \times P(f_3|f_2, f_1) \cdots P(f_m|f_1, f_2 \ldots f_{m-1}) \; , \tag{8}$$

$$\begin{aligned} P(F;C) = P(f_1, \ldots f_m; C) &= P(C)p(f_1|C)P(f_2|f_1, C)P(f_3|f_1, f_2, C) \\ &\times P(f_4|f_1, f_2, f_3, C) \cdots P(f_i|f_1, \ldots, f_m, C) \; . \end{aligned} \tag{9}$$

In our feature selection scheme, the original features are clustered into different groups based on hypergraph clustering and each cluster just includes a small set of features. Therefore, for each cluster, we do not need to use the approximation $\widehat{I}(F;C)$. Instead, we can directly use the multidimensional interaction information $I(F;C)$ criterion for feature selection. Using Parzen windows for probability distribution estimation, we then apply the greedy strategy to select the feature that maximizes the multidimensional mutual information between the features and the output class set. As a result the first feature $f^{'}_{max}$ maximizes $I(f^{'}, C)$, the second selected feature $f^{''}_{max}$ maximizes $I(f^{''}, f^{'}, C)$, the third feature $f^{'''}_{max}$ maximizes $I(f^{'''}, f^{''}, f^{'}, C)$, and so on. For each cluster, we repeat this procedure until $|S| = k$.

4 Experiments and Comparisons

The data sets used to test the performance of our proposed algorithm are the benchmark data sets from the UCI Machine Learning Repository. Table. 1 summarizes the properties of these data-sets. Using the feature selection algorithm outlined above, we make a comparison between our proposed feature selection method (referred to as the HG*plus*MII method) (which utilizes the multidimensional interaction information (MII) criterion and hypergraph clustering for feature selection) and the use of multidimensional interaction information (MII) using the second-order approximation (see Equation (7)).

The experimental results shown in Table. 2 demonstrate that our proposed method (i.e. HG*plus*MII) can achieve higher degree of dimensionality reduction, as it selects a smaller feature subset compared with those obtained using MII with second-order approximation. There are three reasons for this. The first reason is that hypergraph clustering simultaneously considers the information-contribution of each feature and the correlation between features, so the structural information concealed in the data can be effectively identified. The second

Table 1. Summary of UCI benchmark data sets

Data-set	Examples	Features	Classes
Australian	690	14	2
Breast cancer	699	10	2
Pima	768	8	2

reason is that the multidimensional interaction information (MII) criterion is applied to each cluster for feature selection, and can consider the effects of third and higher order dependencies between the features and the class. As a result the optimal feature combination can be located so as to guarantee the optimal feature subset. The third and final reason is that second-order approximation to multidimensional interaction information (MII) simply checks for pair-wise dependencies between features and the class, and so only limited feature subsets can be obtained.

Table 2. The experiment results on three data-sets

Method	Australian	Breast cancer	Pima
MII	$\{f_8,\ f_{14},\ f_5,f_{13}\}$	$\{f_3,\ f_8,\ f_7\}$	$\{f_2,f_8,f_6,f_7\}$
HG$plus$MII	$\{f_8,\ f_9,\ f_5\}$	$\{f_3,\ f_7,\ f_9\}$	$\{f_2,f_6,f_1\}$

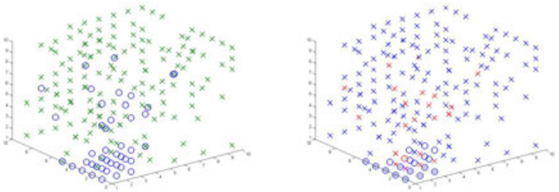
After obtaining the discriminating features, we compute a scatter separability criterion to evaluate the quality of the selected feature subset. This is a well known measure of class separability introduced by Devijiver and Kittler [4], and given by $J(Y) = \frac{|S_w+S_b|}{|S_w|} = \prod_{k=1}^{d}(1+\lambda_k)$, where Y denotes the feature set, λ_k, $k = 1 \ldots d$, are the eigenvalues of matrix $S_w^{-1}S_b$, and S_w and S_b are the between and within class scatter matrices.

Table 3. J value comparisons for two methods on three data sets

Method	Australian	Breast cancer	Pima
MII	2.2832	5.0430	1.3867
HG$plus$MII	2.3010	5.1513	1.3942

In Table. 3, we compare the the performance of the two methods. We find that the effective feature subsets can be obtained using our proposed HG$plus$MII method, e.g., for dataset Australian and Pima, it can achieve a higher discriminability power based on fewer features.This means that our feature selection method can guarantee the optimal feature subset, as it not only achieves higher degree of the dimensionality reduction but also obtains better discriminability power.

After obtaining the discriminating features, we apply a variational EM algorithm to learn Gaussian mixture model on the selected feature subset for the purpose of classification. For the Breast Cancer dataset, we visualize the classification results using the selected feature subset. The classification accuracy achieved using the selected feature subset is 96.3% which is superior to the accuracy of 95.4% achieved by RD-based method [7]. The classification results are shown in Fig. 2. The left hand panel is the data with correct labeling, and the right hand panel is the classification results with the misclassified data highlighted. Because of the unsupervised nature of the variational EM algorithm and



(a) Original data (b) Classification result

Fig. 2. Classification result visualized on 3rd, 7th and 9th features

the Gaussian mixture model, the classification accuracy of 96.3% demonstrates the adequate class separability provided by the selected feature subset.

5 Conclusions

This paper has presented a new graph theoretic approach to feature selection. The proposed feature selection method offers two major advantages. First, hypergraph clustering simultaneously considers the significance of both the features and the correlation between features, and therefor the structural information concealed in the data can be more effectively utilized. Second, the MII criteria takes into account high-order feature interactions with the class, overcoming the problem of overestimated redundancy. As a result the features associated with the greatest amount of joint information can be preserved.

References

1. Balagani, S., Phoha, V.: On the Feature Selection Criterion Based on an Approximation of Multidimensional Mutual Information. IEEE TPAMI 32(7), 1342–1343 (2010)
2. Blum, L., Rivest, L.: Training a 3-Node Neural Network is NP-complete. Neural Networks 5(1), 117–127 (1992)
3. Covões, T., Hruschka, E., de Castro, L., Santos, Á.: A Cluster-based Feature Selection Approach. In: Corchado, E., Wu, X., Oja, E., Herrero, Á., Baruque, B. (eds.) HAIS 2009. LNCS, vol. 5572, pp. 169–176. Springer, Heidelberg (2009)
4. Devijver, A., Kittler, J.: Pattern Recognition: A Statistical Approach, vol. 761. Prentice-Hall, London (1982)
5. Guo, B., Nixon, S.: Gait Feature Subset Selection by Mutual Information. IEEE TSMC, Part A: Systems and Humans 39(1), 36–46 (2008)
6. Kwak, N., Choi, H.: Input Feature Selection by Mutual Information Based on Parzen Window. IEEE TPAMI 24(12), 1667–1671 (2002)
7. Zhang, F., Zhao, Y.J., Fen, J.: Unsupervised Feature Selection based on Feature Relevance. In: ICMLC, vol. 1, pp. 487–492 (2009)

Hypersurface Fitting via Jacobian Nonlinear PCA on Riemannian Space

Jun Fujiki and Shotaro Akaho

National Institute of Advanced Industrial Science and Technology,
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan
{jun-fujiki,s.akaho}@aist.go.jp

Abstract. The subspace fitting method based on usual nonlinear principle component analysis (NLPCA), which minimizes the square distance in feature space, sometimes derives bad estimation because it does not reflect the metric on input space. To alleviate this problem, authors proposed the subspace fitting method based on NLPCA with considering the metric on input space, which is called Jacobian NLPCA. The proposed method is efficient when the metric of input space is defined. The proposed method can be rewritten as kernel method as explained in the paper.

Keywords: fitting, nonlinear principal component analysis, Riemannian space, Euclideanization, kernel.

1 Introduction

Understanding of the structure of data by dimensionality reduction is a fundamental and important task in data processing. To realize this dimensionality reduction, *principal component analysis* (*PCA*) is commonly used. However, PCA can extract only linear (affine) structure of data, and when the data is not assumed to be on a linear space, PCA can not extract appropriate structure of the data. To overcome this drawback, many kinds of *nonlinear PCA* (*NLPCA*) are proposed. The basic idea of NLPCA is that the data which have nonlinear structure is mapped to a high-dimensional space, called feature space, so as to have linear structure in the feature space. Then, original PCA is applied to extract linear structure of the data in the feature space. In the framework of NLPCA, type of extractable nonlinear structure strongly depends on this nonlinear mapping to a feature space, which is called feature mapping. Hence, selecting an appropriate feature mapping is very important to extract appropriate nonlinear structure of the observed data. Generally, type of structure is unknown, while in many applications such as line detection and quadratic curve fitting, type of structure is known and the structure is parameterized by linear parameters. In those cases, NLPCA works very well, however NLPCA sometimes derives a bad estimation. This is because errors of data are measured by the metric in the feature space, not in the input space (the space of observed data) in NLPCA. Since nonlinear mapping does not preserve distances, small errors in the input space sometimes become large in the feature space and large errors in the input space sometimes become small in the feature space, then measuring errors

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 236–243, 2011.

in the feature space is not the best strategy in extracting structure of data. Briefly speaking, usual NLPCA, which is the *least squares* (*LS*) of Euclidean distance in the feature space, is not the *maximum likelihood estimator* (*MLE*) in the input space, and only the MLE in the feature space. Therefore, LS estimator in the input space is investigated, when the input space is an Euclidean space [1,3,6,7,8], and a two-dimensional sphere [5]. In this paper, we extend the framework of these hypersurface fitting methods from an Euclidean space and/or a two-dimensional sphere to a general Riemannian space. The effectiveness of the hypersurface fitting method is shown through experiments.

2 Jacobian NLPCA

When considering hypersurface fitting for the data in m-dimensional input space, the data are mapped into n-dimensional Hilbert space, which is called feature space, so as to have linear structure. By this mapping, hypersurface fitting is resolved into $n-1$-dimensional linear subspace fitting in the feature space. In usual NLPCA, the linear subspace is estimated by minimizing the sum of squared Euclidean distance between data and linear space in the feature space. However, when metric is naturally introduced in the input space, the LS estimator in the feature space sometimes derives bad estimation [2]. Therefore, the approximation of the LS estimator in the input space is proposed when the input space is an Euclidean space [1,3,6,7,8] and/or two-dimensional sphere [5].

2.1 The Metric of Input Space and Feature Space

The input space is assumed to be m-dimensional Riemannian space, and let the Riemannian metric of the space at point $\boldsymbol{x}$ is denoted by $G_{\boldsymbol{x}}$. The observed data in the input space, which are contaminated by noise, is denoted by $\{\boldsymbol{x}_{[d]}\}_{d=1}^{D}$.

In this paper, the metric around the data $\boldsymbol{x}_{[d]}$ is assumed to be approximated by the constant metric $G_{[d]} = G_{\boldsymbol{x}_{[d]}}$, that is, the Riemannian space around the data $\boldsymbol{x}_{[d]}$ is approximated by the tangent affine space at $\boldsymbol{x}_{[d]}$. Under this assumption, the Euclidean distance between the observed data $\boldsymbol{x}_{[d]}$ and the point $\boldsymbol{x} = \boldsymbol{x}_{[d]} + \delta\boldsymbol{x}$, which is close to $\boldsymbol{x}_{[d]}$ is approximated by $r_{[d]} = \sqrt{(\delta\boldsymbol{x})^\top G_{[d]}(\delta\boldsymbol{x})}$.

In hypersurface fitting, the input data $\boldsymbol{x}$ is mapped to the n-dimensional Hilbert space, called feature space, by the feature map $\boldsymbol{\phi} : \boldsymbol{x} \mapsto \boldsymbol{\phi}(\boldsymbol{x})$. Let $J_{\boldsymbol{\phi}}$ be the Jacobian matrix of this mapping, there holds $J_{\boldsymbol{\phi}} = \frac{\partial\boldsymbol{\phi}}{\partial\boldsymbol{x}}$.

By using the value $J_{\boldsymbol{\phi}}$, infinitesimal distance $\delta\boldsymbol{\phi}$ in the feature space is linearly approximated by the infinitesimal distance $\delta\boldsymbol{x}$ in the input space as $\delta\boldsymbol{\phi} = J_{[d]}\delta\boldsymbol{x}$, where $J_{[d]} = J_{\boldsymbol{\phi}_{[d]}}$ is a Jacobian matrix at $\boldsymbol{x}_{[d]}$.

By this approximation, the distance in the input space $r_{[d]}$ is approximated by the quantities in the feature space as

$$r_{[d]} = \sqrt{(\delta\boldsymbol{x})^\top G_{[d]}(\delta\boldsymbol{x})} \approx R_{[d]} = \sqrt{(\delta\boldsymbol{\phi}(\boldsymbol{x}))^\top \mathcal{G}_{[d]}\,\delta\boldsymbol{\phi}(\boldsymbol{x})} \tag{1}$$

where $\mathcal{G}_{[d]} = (J_{[d]}^{+})^\top G_{[d]} J_{[d]}^{+}$ and X^{+} is the Moore-Penrose inverse matrix of X.

2.2 Linear Fitting in the Feature Space

This subsection explains how to fit an $m-1$-dimensional hypersurface (nonlinear subspace) for m-dimensional data which belong to the input space. In this paper, the set of general hypersurface on $\mathcal{R}$ as a family of hypersurfaces described by a linear parameter $\boldsymbol{a}$, that is, the family of fitting curves can be represented as

$$f(\boldsymbol{x};\boldsymbol{a}) = \boldsymbol{a}^\top \boldsymbol{\phi} = 0\,.$$

where $\boldsymbol{a}$ be the parameter which determines a type of fitting curves. For example, the quadratic curve in two-dimensional Euclidean space and the rhumb line in two-dimensional sphere are represented as $\boldsymbol{a}^\top(x^2, xy, y^2, x, y, 1)^\top = 0$ and $\boldsymbol{a}^\top(\sinh^{-1}(\cot\varphi), \psi, 1)^\top = 0$, respectively, where x and y are the first and second components of two-dimensional Euclidean data, and φ and ψ are the colatitude and the longitude of two-dimensional spherical data. When the mapping $\boldsymbol{x} \mapsto \boldsymbol{\phi}$ is considered as a feature mapping, $m-1$-dimensional hypersurface fitting problem is resolved to $n-1$-dimensional linear subspace fitting on feature space. In usual NLPCA, the distance between two points in feature space is measured by Euclidean distance in the feature space, but in the proposed *Jacobian NLPCA* (*JNLPCA*), the distance between two points in feature space is measured by Eq.(1), which is an approximation of the Euclidean distance in the input space. In this framework, the distance between the data and fitting hypersurface is approximated by the quantities in the feature space as following.

Let $\widehat{\boldsymbol{\phi}}_{[d]}$ be the true value of $\boldsymbol{\phi}_{[d]}$ in the feature space, and let $\delta\boldsymbol{\phi}_{[d]} = \widehat{\boldsymbol{\phi}}_{[d]} - \boldsymbol{\phi}_{[d]}$ be the error of each datum. Because the true values are on the fitting hyperplane, there holds $\boldsymbol{a}^\top\widehat{\boldsymbol{\phi}}_{[d]} = 0$, then, $\boldsymbol{a}^\top\delta\boldsymbol{\phi}_{[d]} = -\boldsymbol{a}^\top\boldsymbol{\phi}_{[d]}$ holds. Therefore, the square distance between observed data and the hyperplane is obtained by minimizing $R_{[d]}^2 = (\delta\boldsymbol{\phi}(\boldsymbol{x}))^\top\mathcal{G}_{[d]}\,\delta\boldsymbol{\phi}(\boldsymbol{x})$ under the condition $\left|\boldsymbol{a}^\top\delta\boldsymbol{\phi}_{[d]}\right|^2 = \left|\boldsymbol{a}^\top\boldsymbol{\phi}_{[d]}\right|^2$.

This minimization problem can be solved by Cauchy-Schwarz inequality as $\min R_{[d]}^2 = \dfrac{\boldsymbol{a}^\top\left[\boldsymbol{\phi}_{[d]}\boldsymbol{\phi}_{[d]}^\top\right]\boldsymbol{a}}{\boldsymbol{a}^\top\mathcal{G}_{[d]}^+\boldsymbol{a}}$, and the sum of square distance between observed data and hypersurface is approximated by

$$\mathcal{E}(\boldsymbol{a}) = \sum_{d=1}^{D}\frac{\boldsymbol{a}^\top\left[\boldsymbol{\phi}_{[d]}\boldsymbol{\phi}_{[d]}^\top\right]\boldsymbol{a}}{\boldsymbol{a}^\top\mathcal{G}_{[d]}^+\boldsymbol{a}}\,, \tag{2}$$

which is the sum of Rayleigh quotients. Then, JNLPCA estimates hypersurface by minimizing $\mathcal{E}(\boldsymbol{a})$.

2.3 Algorithm

In this section, the algorithm to minimize Eq.(2) [1] is introduced. When the approximation of $\boldsymbol{a}$ is computed as $\widehat{\boldsymbol{a}}$, the value $\boldsymbol{a}^\top\mathcal{G}_{[d]}^+\boldsymbol{a}$ is fixed as $\mu_{[d]} = \widehat{\boldsymbol{a}}^\top\mathcal{G}_{[d]}^+\widehat{\boldsymbol{a}}$, and Eq.(2) is approximated by a quadratic form as

$$\mathcal{E}(\boldsymbol{a}) \approx \mathcal{E}'(\boldsymbol{a}) = \sum_{d=1}^{D}\boldsymbol{a}^\top\left[\mu_{[d]}^{-1}\boldsymbol{\phi}_{[d]}\boldsymbol{\phi}_{[d]}^\top\right]\boldsymbol{a} = \boldsymbol{a}^\top\mathcal{F}\Lambda^{-1}\mathcal{F}\boldsymbol{a}$$

where $\Lambda = \texttt{diag}\left\{\mu_{[1]}, \ldots, \mu_{[D]}\right\}$ and $\mathcal{F} = (\boldsymbol{\phi}_{[1]}, \ldots, \boldsymbol{\phi}_{[D]})$. Under this approximation, $\boldsymbol{a}$ is approximated by the unit eigen vector of $\mathcal{F}\Lambda^{-1}\mathcal{F}$ corresponding to the minimum eigen value, which is denoted by $\texttt{UnitMinEigenVec}\left[\mathcal{F}\Lambda^{-1}\mathcal{F}\right]$.

By using this, the iterative algorithm to compute $\boldsymbol{a}$ is presented. In the following algorithm, upper right suffix $[k]$ represents the values in the k-th step. Especially, upper right suffix of initial values are $[0]$. How to compute the initial values are introduced later.

(1) Compute initial values $\{\mu_{[d]}^{[0]}\}_{d=1}^{D}$.
(2) Repeat (a) and (b) till converge:
  (a) $\widehat{\boldsymbol{a}}^{[k+1]} = \texttt{UnitMinEigenVec}[\mathcal{F}(\Lambda^{[k]})^{-1}\mathcal{F}]$
  (b) $\mu_{[d]}^{[k+1]} = (\widehat{\boldsymbol{a}}^{[k+1]})^{\top}\mathcal{G}_{[d]}^{+}(\widehat{\boldsymbol{a}}^{[k+1]})$.

2.4 Euclideanization of Metric as an Initial Value

In this section, we introduced zeroth (0th) order Euclideanization of metric as a method to compute an initial value of the proposed method. Akaho [1] uses the LS estimator in feature space (result of usual NLPCA) as an initial value. The usual LS does not consider the change of metric, that is, estimate hypersurface parameters under the assumption of the metric of feature space is the identity matrix. In this case, the initial value of $\boldsymbol{a}$ is derived by setting $\{\mu_{[d]}^{[0]} = 1\}_{d=1}^{D}$. Then the energy function (Eq.2) is approximated as $\mathcal{E}(\boldsymbol{a}) \approx \boldsymbol{a}^{\top}\left[\mathcal{F}\mathcal{F}^{\top}\right]\boldsymbol{a}$, and the initial value of the estimation is set to $\widehat{\boldsymbol{a}}^{[1]} = \texttt{UnitMinEigenVec}\left[\mathcal{F}\mathcal{F}^{\top}\right]$, which is the LS estimator. The usual LS estimator is a good initial value for JNLPCA, but sometimes the usual LS estimator estimates bad parameters. Therefore, Euclideanization of metric on hypersphere is proposed to estimate small hypersphere [4]. This paper extends the Euclideanization for general Riemannian spaces. The concept of Euclideanization is the adjustment of the metric of feature space to keep the metric of the input space. In this sense, the method proposed in previous section is also a kind of Euclideanization. To distinguish these Euclideanizations, Euclideanization proposed by [4] is called zeroth order Euclideanization, and Euclideanization proposed in this paper is called first order Euclideanization. Zeroth order Euclideanization of metric is first introduced as the adjustment of the length by using the enlargement of volume element. The concept of zeroth order Euclideanization is when m-dimensional volume is enlarged k times by feature mapping, length is expected to be enlarged $k^{\frac{1}{m}}$ times. Therefore, LS in input space is approximated by weighted LS in feature space, and the weights are computed by $k^{-\frac{2}{m}}$. Because the Jacobian matrix of the mapping $\boldsymbol{x} \mapsto \boldsymbol{\phi}$ satisfies $\delta\boldsymbol{\phi} = J_{\boldsymbol{\phi}}\delta\boldsymbol{x}$, the enlargement of m-dimensional volume element by the mapping around $\boldsymbol{x}_{[d]}$ is $(\texttt{Det}\,\mathcal{G}_{[d]})^{-\frac{1}{2}}$, where $\texttt{Det}\,\mathcal{G}_{[d]} = \dfrac{\texttt{det}\,G_{[d]}}{\texttt{det}\left\{J_{[d]}^{\top}J_{[d]}\right\}}$. Therefore, the one-dimensional length is expected to enlarged $(\texttt{Det}\,\mathcal{G}_{[d]})^{-\frac{1}{2m}}$ times, and LS in input space is approximated by weighted LS in feature space, and the weights are computed by $(\texttt{Det}\,\mathcal{G})^{\frac{1}{m}}$.

By the Euclideanization of the metric, $\mathcal{E}(\boldsymbol{a})$ is approximated as $\boldsymbol{a}^\top \left(\mathcal{F} D^{\frac{1}{m}} \mathcal{F}\right) \boldsymbol{a}$ where $D = \mathrm{diag}\left\{\mathtt{Det}\,\mathcal{G}_{[1]}, \ldots, \mathtt{Det}\,\mathcal{G}_{[D]}\right\}$. Then, the initial value is computed as $\widehat{\boldsymbol{a}}^{[0]} = \mathtt{UnitMinEigenVec}[\mathcal{F} D^{\frac{1}{m}} \mathcal{F}]$. Note that the change of infinitesimal distance by the mapping is as already mentioned as $r^2_{[d]} \approx R^2_{[d]} = \delta\boldsymbol{\phi}^\top \mathcal{G}_{[d]} \delta\boldsymbol{\phi}$, the zeroth order Euaclideanization is easy to compute by considering the change of infinitesimal distance. Now, we compare with three types of distance in the feature space (Table 1). In the three types of distance, numerator is common, and denominator is different. For usual LS, the metric in the feature space is the identity matrix. For the zeroth order Euclideanization, the metric is a scalar matrix. For the first order Euclideanization, the metric is a symmetric matrix. Here, the zeroth order Euclideanization is an approximation of that of the first order as $\mathcal{G}^+_{[d]} \approx (\mathtt{Det}\,\mathcal{G}^+_{[d]})^{\frac{1}{m}}\, \mathtt{I}_n$ on m-dimensional space.

Table 1. Three types of distance

	ordinary LS	0th order	1st order (JNLPCA)
$r^2_{[d]}$	$\dfrac{\{f(\boldsymbol{x}_{[d]}; \boldsymbol{a})\}^2}{\boldsymbol{a}^\top\, \mathtt{I}_n\, \boldsymbol{a}}$	$\dfrac{\{f(\boldsymbol{x}_{[d]}; \boldsymbol{a})\}^2}{\boldsymbol{a}^\top \left\{\left(\mathtt{det}\,\mathcal{G}^+_{[d]}\right)^{\frac{1}{2}}\, \mathtt{I}_n\right\} \boldsymbol{a}}$	$\dfrac{\{f(\boldsymbol{x}_{[d]}; \boldsymbol{a})\}^2}{\boldsymbol{a}^\top\, \mathcal{G}^+_{[d]}\, \boldsymbol{a}}$

3 Quadratic Curve Fitting on Plane

In this section, an experimental result to show the effectiveness of JNLPCA is provided. We generate 50-data from $y = x^2$ of its x-coordinate is chosen uniformly from the interval $[-3, 3]$ and adding the Gauss noise of 0.04-standard deviation for each coordinate. Figure 1 shows the fitting result when feature mapping is chosen as $\boldsymbol{\phi}(\boldsymbol{x}) = \left(x^2\ 2xy\ y^2\ 2x\ 2y\ 1\right)^\top$, which is denoted as DLT (direct linear transformation). For DLT, JNLPCA derives better result than NLPCA. Figure 2 shows the effect of selecting feature mapping for NLPCA and JNLPCA. The horizontal axis shows the fitting error of DLT, and the vertical axis shows the fitting error of polynomial kernel mapping $\boldsymbol{\phi}(\boldsymbol{x}) = \left(x^2\ \sqrt{2}xy\ y^2\ \sqrt{2}x\ \sqrt{2}y\ 1\right)^\top$.

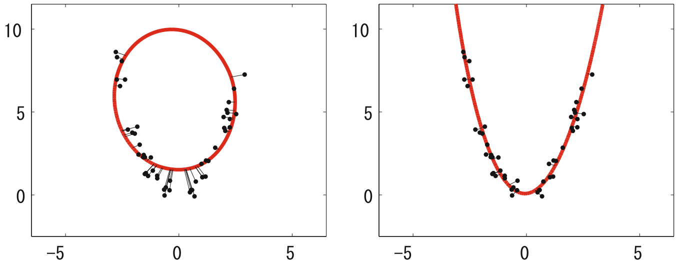


Fig. 1. Fitting results: NLPCA (left), JNLPCA (right)

The left of Fig. 2 is for NLPCA, and the right of Fig. 2 is for JNLPCA. When the estimation by using DLT and polynomial kernel mapping is the same, the points depicts in Fig. 2 is lying on the diagonal line. From Fig. 2, estimation by NLPCA is sensitive to selecting the feature mapping but JNLPCA is not sensitive to the choice of the feature mapping. This is because JNLPCA reflects the distance in the input space, but NLPCA does not reflect the distance in the input space. Figure 3 plots the true distance versus the approximated distance by JNLPCA. The left of Fig. 3 shows the mean of square distance of the approximated distance and the true distance in the input space of each iteration, and the right of Fig. 3 shows the distance of the approximated distance and the true distance in the input space of each datum after convergence. Both figures show the approximated distance approximates true distance very well. Figure 4 plots the approximated distance by JNLPCA versus the approximated distance by zeroth order Euclideanization. The figure shows zeroth order Euclideanization is very good initial value for JNLPCA.

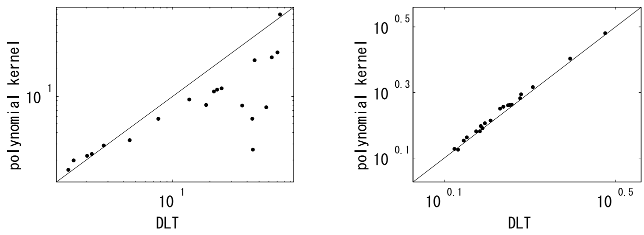


Fig. 2. Distance between data and curve: NLPCA (left), JNLPCA (right)

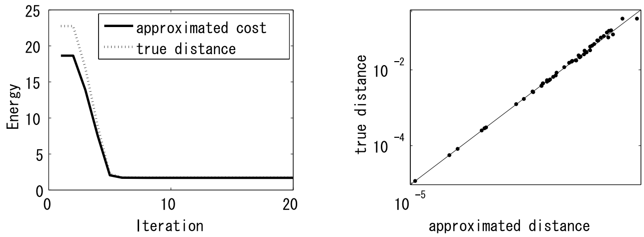


Fig. 3. Approximated distance and true distance by JNLPCA: The number of iteration vs. energy function (left) and comparison between true distance and approximated distance by JNLPCA (right)

4 Rhumb Line Fitting on 2-Dimensional Sphere

In this section, JNLPCA is applied for the estimation of rhumb line on S^2 data by projecting into Mercator projection plane. In the experiment, the true rhumb

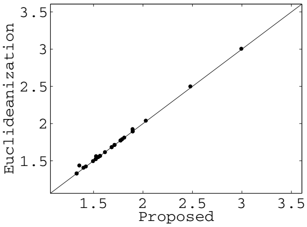


Fig. 4. Approximated distance of JNLPCA (horizontal axis) vs. Euclideanization

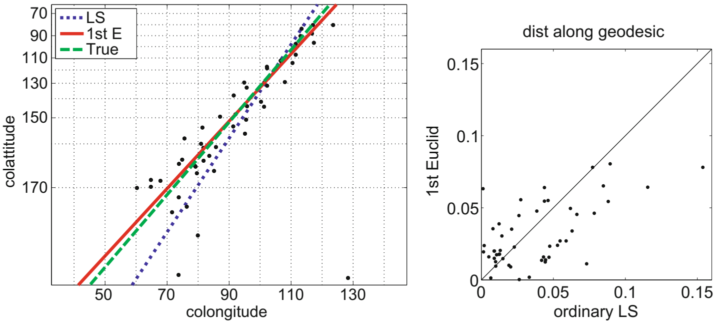


Fig. 5. Rhumb line fitting: Fitting result (left) and error distribution of rhumb line fitting (right; horizontal axis is ordinary LS error and vertical axis is first order Euclideanization error)

line path through $\phi = (70, 120)^\top$[deg] and $(175, 60)^\top$[deg]. We generate 50-data from uniform distribution on the arc of the rhumb line and adding the Gauss noise of 0.05-degree standard deviation for each of 3D coordinate of S^2 data and normalized as normal vector. Left of Fig. 5 is the fitting result of rhumb line. Green dashed line is the ground truth, blue dotted line is the estimation by ordinary LS, and red solid line is the estimation by first order Euclideanization. From left of Fig. 5, Euclideanization works very well. Right of Fig. 5 shows the error distribution. From right of Fig. 5, errors for each data tend to be small by using Euclideanization.

5 Conclusion

In this paper, we proposed a hypersphere fitting method via JNLPCA, which minimizes the sum of the squares approximated Euclidean distance in the input space. We also extended the zeroth Euclideanization of the metric from hyperspherical data to a general Riemannian spaces. Experiments showed the effectiveness of JNLPCA. The fitting method has possibilities to apply for many kinds of problems which can be regarded as nonlinear dimension reductions.

The proposed method can be extended to the kernel method by defining a kernel function $k(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\phi}(\boldsymbol{y})$ and its diffential $\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\partial k(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{x}^\top} = J_{\boldsymbol{\phi}}(\boldsymbol{x})^\top \boldsymbol{\phi}(\boldsymbol{y})$, which is named Jacobian kernel. Although it is not proved that the representer theorem holds or not, suppose that $\boldsymbol{a}$ can be represented by a linear combination of $\boldsymbol{\phi}_{[d]}$ as $\boldsymbol{a} = \sum_p \alpha_{[d]} \boldsymbol{\phi}_{[d]} = \boldsymbol{\Phi} \boldsymbol{\alpha}$. Let $\mathcal{K}$ and $\boldsymbol{\mathcal{K}}_{[d]}$ be defined as

$$(\mathcal{K})_{ij} = k(\boldsymbol{x}_{[i]}, \boldsymbol{x}_{[j]}), \quad \mathcal{K} = \left(\mathcal{K}_{[1]} \cdots \mathcal{K}_{[D]}\right),$$
$$\boldsymbol{k}_{[i][j]} = \boldsymbol{k}(\boldsymbol{x}_{[i]}, \boldsymbol{x}_{[j]}), \quad \boldsymbol{\mathcal{K}}_{[d]} = \left(\boldsymbol{k}_{[d][1]} \cdots \boldsymbol{k}_{[d][D]}\right)^\top,$$

respectively, there hold $\mathcal{K}_{[d]} = \boldsymbol{\Phi}^\top \boldsymbol{\phi}_{[d]}$ and $\boldsymbol{\mathcal{K}}_{[d]} = \boldsymbol{\Phi}^\top J_{[d]}$, and therefore, $\boldsymbol{a}^\top \left[\boldsymbol{\phi}_{[d]} \boldsymbol{\phi}_{[d]}^\top\right] \boldsymbol{a} = \boldsymbol{\alpha}^\top \mathcal{K}_{[d]} \mathcal{K}_{[d]}^\top \boldsymbol{\alpha}$ and $\boldsymbol{a}^\top \mathcal{G}_{[d]}^{+} \boldsymbol{a} = \boldsymbol{\alpha}^\top \boldsymbol{\mathcal{K}}_{[d]} G_{[d]}^{-1} \boldsymbol{\mathcal{K}}_{[d]}^\top \boldsymbol{\alpha}$ hold. Then Eq.(2) can be rewritten as

$$\mathcal{E}_{\mathrm{kernel}}(\boldsymbol{\alpha}) = \sum_{d=1}^{D} \frac{\boldsymbol{\alpha}^\top \mathcal{K}_{[d]} \mathcal{K}_{[d]}^\top \boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top \boldsymbol{\mathcal{K}}_{[d]} G_{[d]}^{-1} \boldsymbol{\mathcal{K}}_{[d]}^\top \boldsymbol{\alpha}} \tag{3}$$

which is also the sum of Rayleigh quotients, and estimation is realized by minimizing $\mathcal{E}_{\mathrm{kernel}}(\boldsymbol{\alpha})$ in the same way as the proposed method.

References

1. Akaho, S.: Curve fitting that minimizes the mean square of perpendicular distances from sample points. In: SPIE, Vision Geometry II (1993)
2. Akaho, S.: SVM that maximizes the margin in the input space. In: Shannon, S. (ed.) Artificial Intelligence and Computer Science, ch. 5, pp. 139–154 (2005)
3. Chojnacki, W., Brooks, M.J., van den Hangel, A., Gawley, D.: On the fitting of surface to data with covariances. IEEE TPAMI 22(11), 1294–1303 (2000)
4. Fujiki, J., Akaho, S.: Small hypersphere fitting by Spherical Least Square. In: ICONIP 2005, pp. 439–444 (2005)
5. Fujiki, J., Akaho, S.: Curve fitting by spherical least squares on two-dimensional sphere. Subspace 2009 Workshop in Conjunction with ICCV 2009 (2009)
6. Kanatani, K., Sugaya, Y.: Unified computation of strict maximum likelihood for geometric fitting. Journal of Math. Imaging and Vision 38(1), 1–13 (2010)
7. Sampson, P.D.: Fitting conic sections to very scattered data: an iterative refinement of the Bookstein algorithm. Comput. Vision, Graphics, and Image Processing 18, 97–108 (1982)
8. Taubin, G.: Estimation of planar curves, surfaces, and nonplanar space curves defined by implicit equations with applicatons to edge and range image segmentation. IEEE TPAMI 13(11), 1115–1138 (1991)

A Robust Approach to Multi-feature Based Mesh Segmentation Using Adaptive Density Estimation

Tilman Wekel and Olaf Hellwich

Computer Vision and Remote Sensing, TU-Berlin
{t.wekel,olaf.hellwich}@tu-berlin.de

Abstract. In this paper, a new and robust approach to mesh segmentation is presented. There are various algorithms which deliver satisfying results on clean 3D models. However, many reverse-engineering applications in computer vision such as 3D reconstruction produce extremely noisy or even incomplete data. The presented segmentation algorithm copes with this challenge by a robust semi-global clustering scheme and a cost-function that is based on a probabilistic model. Vision based reconstruction methods are able to generate colored meshes and it is shown, how the vertex color can be used as a supportive feature. A probabilistic framework allows the algorithm to be easily extended by other user defined features. The segmentation scheme is a local iterative optimization embedded in a hierarchical clustering technique. The presented method has been successfully tested on various real world examples.

Keywords: mesh segmentation, density estimation, hierarchical clustering, variational shape approximation.

1 Introduction

3D reconstruction is an important research field in computer vision. Recent advances in software and hardware technology allow to acquire highly complex and detailed volumes based on common, heterogeneous images. 3D reconstruction is a reverse-engineering discipline and the acquired data is unstructured, noisy or incomplete. The outcome of a laser-scanner or a vision based reconstruction is a point-cloud or a polyhedral surface without semantic information [1]. In order to enable an efficient representation and further use such as navigation, computer aided design or structural analysis, high level information needs to be obtained from the data. Mesh segmentation refers to the process of subdividing a polyhedral surface into several segments in order to obtain a more suitable representation or to guide further processing algorithms. In this work, a segmentation that maximizes a given probability function is defined as optimal. While many state-of-the-art algorithms deliver satisfying results on clean 3D models, there is still a lack of attention to the segmentation of real world data which is far more challenging. The design of the segmentation algorithm needs to be

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 244–252, 2011.

adapted in order to cope with typical deficits of real-world data. In this paper, a novel approach to the segmentation of noisy polyhedral surfaces is presented. The introduced segmentation scheme is inspired by the well known Expectation-Maximization-Algorithm (EM) [2] and the hierarchical clustering (HC) [3]. Underpinned by a probabilistic model, a cost function is introduced that is based on an adaptive probability density estimation. The theoretical model allows to incorporate multiple local features such as color, normal vectors or curvature. The paper is organized as follows. Related state of the art approaches are briefly reviewed in the following section. In Section 3, the probabilistic model and the probability density estimation is derived. The design of the incorporated features is presented in Section 4. The actual algorithm is presented in Section 5, while experimental results are given in Section 6. Finally, an outlook as well as possible applications are shown in the last section.

2 Related Work

Hierarchical clustering is an unsupervised learning technique and has been studied for decades [4]. The application to mesh segmentation is also not new. The bottom-up version initializes each face with its own cluster. The segmentation is now performed by an iterative merge of adjacent clusters according to a cost function. The segmentation scheme presented by Garland et al. provides a key component for this work [5]. It uses a dual graph to describe the current cluster configuration. Each node in this graph represents a cluster and adjacent clusters are connected by edges. A quadric-error based metric is used to estimate the merging costs. They introduce a compactness heuristic to avoid irregular shapes. Attene et al. present an advanced cost function for this algorithm, where basic geometric models such as planes, spheres and cylinders are fitted to a potential cluster pair in order to estimate the costs [6]. The algorithm is applied to a selection of clean 3D models. Iterative clustering also has its origin in the field of unsupervised learning. Related mesh segmentation schemes mostly follow a k-means strategy, also known as Lloyd's algorithm. The number of clusters must be given in advance and each cluster is represented by a center, sometimes called proxy [7]. Each facet is assigned to one cluster before its proxy is re-fitted to the current partitioning. The two step procedure is repeated until convergence. The algorithm can run into local minima if the initial cluster centers are not estimated appropriately. A segmentation algorithm that is based on Lloyd's algorithm is presented by Shlafman et al. [8]. The cost function incorporates the angular as well as the Euclidean distance between face and cluster. In contrast to this work, the algorithm is initialized with one cluster and new clusters are added one after another until the desired number of clusters is reached. The variational shape approximation (VSA) algorithm presented by Cohen-Steiner et al. uses a modified Lloyd scheme, but in the context of mesh simplification [7]. The cost function is based on the normal vectors. A region growing scheme is used to assign the facets in order to ensure connected clusters. There are many papers that suggest improvements and extensions for this algorithm. However, the cluster initialization problem remains a crucial aspect of this approach. Wu and Kobbelt use geometric models such as cylinders, spheres and planes as proxies. It allows to segment heterogeneous geometry on the expense of the computational complexity [9]. Julius et al. present a slightly modified version of the VSA-algorithm, that partitions the surface into developable charts [10]. Chiosa and Kolb present a combination of the VSA- and the HC-algorithm. Compared to our work, they use a simple normal-vector based cost function and each facet is initialized with its own cluster [11].

3 Propabilistic Model and Density Estimation

In this section, a probabilistic model is derived in order to formulate the given problem precisely. In a clustering analysis, one tries to find an assignment of a set of observations to clusters $C = \{c_0, \ldots, c_e\}$, according to the similarity of specific properties. The observations are the facets $D = \{d_0, \ldots, d_n\}$ of a polyhedral surface $S = \{D\}$. Note, that d_i does not only contain the geometric representation of the respective facet, but also features such as color, normal vectors or other local properties. Assuming that the observations D are a sample drawn from a mixed model $\theta = \{\theta_0, \ldots, \theta_e\}$, one tries to find a $\hat{\theta}$, that maximizes the likelihood. A set of hidden variables $Z = \{z_0, \cdots, z_n\}$ is introduced, which represents the assignment for each facet d_i to the individual model components of θ. The resulting problem can be stated as:

$$\hat{\theta}(D) = \hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} p(d_i, z_i \mid \theta) \ . \tag{1}$$

Unfortunately, there is no closed form solution to this problem and neither the model parameters nor the hidden variables are known. However, if fixed model parameters θ are assumed, an assignment $\hat{Z}$ can be computed in the t^{th}-iteration in order to maximize:

$$\hat{Z}^{(t)} = \arg\max_{Z} \prod_{i=1}^{n} p(z_i \mid d_i,\ \theta^{(t)}) \ . \tag{2}$$

Here, each facet is assigned to a cluster according to:

$$z_i = \arg\max_{m \in [0,e]} p(d_i \ |\theta_m) \ . \tag{3}$$

Note, that this describes a hard class assignment, each facet is assigned to one cluster only. Given $\hat{Z}^{(t)}$ in the t^{th}-iteration, the new model parameters can be estimated using maximum likelihood-estimation:

$$\hat{\theta}^{(t+1)} = \arg\max_{\theta} \prod_{i=1}^{n} p({z_i}^{(t)} \mid \theta) \ , \tag{4}$$

where $p({z_i}^{(t)} \mid \theta^{(t)})$ is computed according to a probability density function. The presented derivation leads to Lloyd's algorithm, also known as k-means, which tries to find optimal clusters by iterating these two steps until convergence [12]. Another approach to that problem is a greedy-algorithm that is called hierarchical clustering [3]. The bottom-up approach initializes each facet d_i with its own cluster c_j and iteratively merges the cluster pairs according to a cost-function. Here, the probability density of a cluster merge is approximated by:

$$p(c_n = c_a \cup c_b) \approx \prod_{d_i \in c_a} p(d_i|\ \theta_b) \cdot \prod_{d_i \in c_b} p(d_i|\ \theta_a)\ , \tag{5}$$

which allows to express the compatibility of two clusters without estimating a new parameter vector θ_n. Each cluster c_j is represented by a model and its corresponding parameter vector θ_j. It is now shown, how a kernel based density estimator can be used to quantify θ_j and the likelihood function $p(d|\theta_j)$. In many segmentation algorithms each cluster is only represented by a single feature vector [7]. From a probabilistic perspective, this implicitly assumes that the features within one cluster are samples, drawn from a Gaussian distribution. Non-parametric kernel based methods are very suitable if no prior knowledge is given [13]. The goal of a density estimation is to find a function $\hat{p}_j$, that approximates the true probability density function $p(d|\theta_j)$ as good as possible. Now, consider that θ_j is updated in the maximization step based on all facets currently assigned to c_j. The kernel density estimator turns out to be:

$$p(d|\theta_j) \approx \hat{p}_j(d) = \frac{1}{\sum\limits_{i=1}^{|c_j|} w_i} \sum_{i=1}^{|c_j|} \frac{w_i}{h_i} K(\frac{d - d_i}{h_i}),\ d_i \in c_j, \tag{6}$$

where an exponential kernel function is used for K. h_i is the bandwidth which controls the smoothness of a given kernel function. An adaptive bandwidth allows to model fine details in densely observed areas while reducing the variance in areas with only a few data points. According to the standard adaptive two-step stage estimator presented by Abramson [14], h_i can be approximated by a function $\hat{p}'_j$ which is again a density estimator, but with $\hat{h} = 1$:

$$h_i = (\bar{g_j} / \hat{p}'_j(d_i))^{0.5}\ , \tag{7}$$

where $\bar{g_j}$ is the geometric mean over all $\hat{p}'_j(d_i),\ d_i \in c_j$. As mentioned before, d_i is a multidimensional vector which contains several features such as color or normal vectors. The quality of the probability function depends on the sample density which decreases rapidly with higher dimensions. However, the vector d_i can be decomposed into low-dimensional sub-vectors that are assumed to be statistically independent $d_i = ({d_i}^{(1)}, \cdots, {d_i}^{(m)})^T$, where m is the number of independent feature vectors. Each sub-vector contains the parameters of a feature such as normal vectors or color. The relatively small dimension of the sub-vectors allows to perform a seperate and reliable density estimation for each ${d_i}^{(k)}$. Accordingly, the likelihood can then be calculated as: $p(d_i|\theta_j) = \prod\limits_{k=1}^{m} p({d_i}^{(k)}|\theta_j)$.

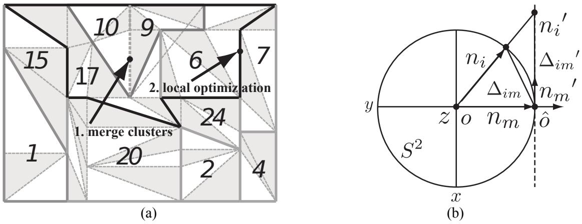


Fig. 1. Segmentation scheme and mapping of normal vectors

4 Features

It is shown by Garland et al., that the normal vectors can be used as reasonable segmentation features [5]. Alternatively, complex model fitting approaches are quite successful but assume a specific geometric structure such as spheres or cylinders and the computation costs are relatively high. The space of all normal vectors is a 2D manifold embedded in $\mathbb{R}^3$. A mapping $L : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ is derived in the following. The Gauss map projects each point p_i of a surface to the unit sphere $S^2 = \{(x, y, z) \in \mathbb{R}^3 | x^2 + y^2 + z^2 = 1\}$ by centering the corresponding normal vector n_i at the origin of S^2, where $\|n_i\| = 1$. The surface of the unit sphere is a $2D$ manifold embedded in $\mathbb{R}^3$. If the orientation ambiguity is taken into account, one hemisphere is sufficient to represent all possible normal vectors. Instead of doing calculus in spheric coordinates, the vector n_i is projected to a point $n_i{}'$ on a tangent plane, as shown in (b) of fig. 1. The tangent plane is centered at n_m and orthogonal to n_m. n_m is the mean normal vector of a cluster. The two-dimendional difference vector Δ'_{im} now represents the feature vector $d_i{}^{(0)}$. If color information is available, for example in the context of vision based reconstruction, it might be used as an additional feature. It is shown, that the HSI - model is more compatible to human vision [15]. The hue and saturation value is very discriminative. The intensity channel is not taken into account. The color information for each facet is represented by $d_i{}^{(1)}$. Although the features presented in this section are quite promising, others might be incorporated as well in order to customize the segmentation algorithm according to the specific application.

5 Segmentation Scheme

The probabilistic framework that is presented in Section 3 is now used to derive a mesh segmentation algorithm that represents a combination of both unsupervised learning methods. The segmentation scheme consists of an outer algorithm, which is a slightly modified hierarchical clustering and an embedded Lloyd's algorithm which optimizes the modified clusters after each merging. Consider a polyhedral surface $S = \{D\}$, that consists of facets $D = \{d_0, \ldots, d_n\}$. The goal is a suitable partitioning of the data into e non overlapping and connected subsets of facets, $C = \{c_0, \ldots, c_e\}$. According to Garland et al. , a cluster graph $G = \{C, A\}$ is introduced, where the nodes C represent the clusters and the edges A define the neighborhood in between the clusters [5]. Only adjacent clusters are considered for merging. A merge is performed by a collapse of an edge in the clustering graph. Each edge is assigned with a merging probability according to equation 5. The set of all collapses is efficiently managed in a heap sorted by costs. Similar to the approach presented by Chiosa and Kolb, the algorithm is interrupted after each iteration and an iterative optimization scheme is applied [11]. In contrast to their work, this algorithm only considers the local neighborhood consisting of the merged cluster itself and its adjacent clusters. Lloyd's algorithm is embedded in the hierarchical clustering scheme and a local subset of clusters is optimized after merging as it can be seen in (a) of fig. 1. The optimization is done by a repeated expectation and maximization step. Given a fixed number of s clusters in the local neighborhood, s seed facets are picked at random to initialize them. The seed facets are paired with the corresponding cluster labels and pushed into a priority queue. At each iteration, a facet $d_{(q,j)}$ is extracted from the queue. If not already assigned to a cluster, all adjacent facets of $d_{(q,j)}$ are tested against the $j-th$ cluster by calculating the probability of the facet, given θ_j according to equation 3. If not already contained, the facets and their assignment probabilities are then inserted into the queue. Please note that this ensures fully connected clusters. The expectation step is finished when the priority queue is empty. In a second step, the model parameters of each cluster is recalculated using the density estimation in order to maximize the term in equation 4. Now for each cluster, the facet with highest assignment probability density is used as the new seed face. The inner algorithm terminates either after a user defined number of maximum iterations or if a convergence threshold is reached and the hierarchical clustering scheme continues. Originally, the hierarchical clustering algorithm starts with initializing a cluster for each facet. However, the desired number of clusters c_e is typically significantly smaller than the number of facets n and the algorithm is bootstrapped with c_s clusters, where $c_e < c_s < n$. The initial c_s clusters are optimized by Lloyd's algorithm before the actual segmentation starts.

6 Evaluation

In the following, the algorithm is tested on real-world data. A detailed evaluation of the segmentation result is a challenging issue by itself and beyond the focus of that paper. The result highly depends on the applied cost function which always assumes a specific application. Mostly, results are compared to human segmentation. Furthermore, the available model databases consist of clean and uncolored meshes which are not suitable for this approach [16]. Consider the polyhedral surfaces presented in (a) and (b) of fig. 2. The data has been obtained by a multiple vision reconstruction tool-chain. The design of the incorporated features highly depends on the context. Consequently, the evaluation focuses on

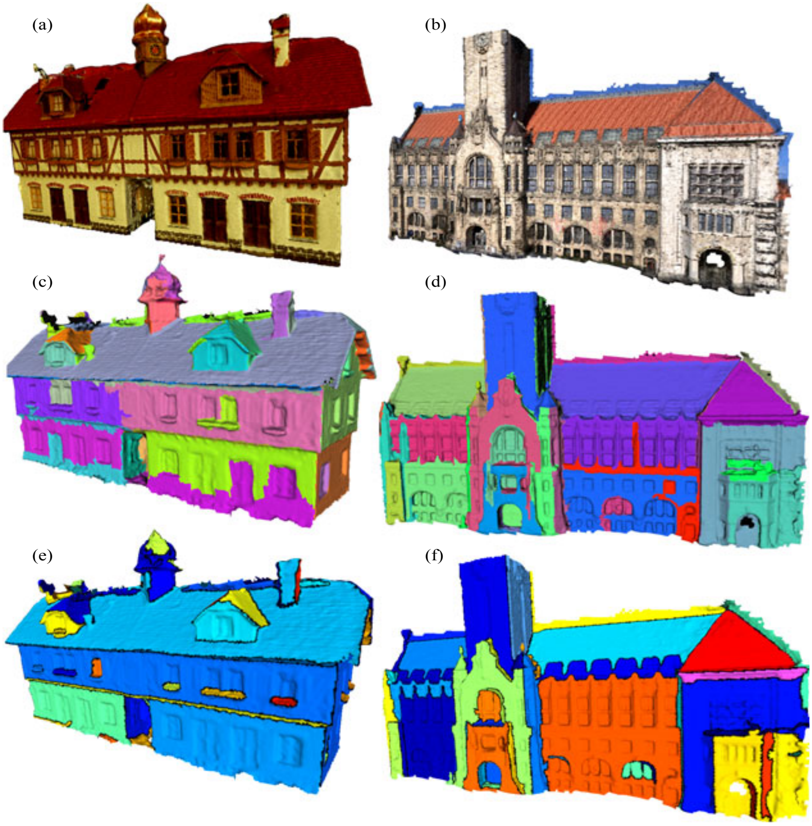


Fig. 2. Two examples are investigated, where (a) represents a plastic house and (b) is the reconstruction of a town hall. The result of a related hierarchical clustering algorithm can be seen in (c) and (b). The segmentation achieved by the presented approach is given in (e) and (f).

colored meshes with mostly planar shapes. The result of a normal vector based hierarchical clustering approach can be seen in (c) and (d). The segmentation has been computed by the tool that comes with the corresponding paper of Attene et. al [6]. Although the algorithm produces a reasonable segmentation that respects the general geometry, the quality of the cluster shapes as well as the partitioning is not satisfying. The result of the presented algorithm can be seen in (e) and (f). The shape of the clusters is not unnecessarily complex and the boundary looks smoother. The distribution of the clusters appears to be more reasonable. Please note that the propabilities of the independent features are multiplied and one cluster should be characterized by both similar color and similar normal vectors. In regions characterized by homogeneous normal vectors such as the roof of the town hall (b), the color becomes the dominant feature, as it can be seen in (f). It turns out, that the local Lloyd optimization does not need to be performed after every single merge. Instead, computation time can be reduced by executing multiple merges before the respective region is optimized.

7 Conclusion

The proposed algorithm has been shown to be successful by various examples. This work holds several new contributions. A probabilistic model is used to introduce a combination of hierarchical and iterative clustering and the incorporation of multiple features. The cost function is based on an adaptive probability density function in order to avoid strong assumptions. The color is presented as an useful feature especially in the context of vision-based surface reconstruction. Furthermore, it is shown, how the normal vectors can be efficiently represented and compared in 2D space. Although the presented approach is promising, the general problem of mesh segmentation remains challenging. In future work, the performance as well as the reliability of the presented algorithm will be investigated.

References

1. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from internet photo collections. Int. J. Comput. Vision 80, 189–210 (2008)
2. Moon, T.K.: The expectation-maximization algorithm. IEEE Signal Processing Magazine 13(6), 47–60 (1996)
3. Heller, K., Ghahramani, Z.: Bayesian hierarchical clustering. In: Proceedings of the 22nd International Conference on Machine Learning, pp. 297–304. ACM, New York (2005)
4. Anderberg, M.R.: Cluster Analysis for Applications. Ac. Press, New York (1973)
5. Garland, M., Willmott, A., Heckbert, P.S.: Hierarchical face clustering on polygonal surfaces. In: Proceedings of the 2001 Symposium on Interactive 3D Graphics, I3D 2001, pp. 49–58. ACM, NY (2001)
6. Attene, M., Falcidieno, B., Spagnuolo, M.: Hierarchical mesh segmentation based on fitting primitives. Vis. Comput. 22, 181–193 (2006)
7. Cohen-Steiner, D., Alliez, P., Desbrun, M.: Variational shape approximation. In: ACM SIGGRAPH 2004 Papers (2004)
8. Shlafman, S., Tal, A., Katz, S.: Metamorphosis of polyhedral surfaces using decomposition. In: CG Forum, vol. 21, pp. 219–228. Wiley Online Library (2002)
9. Wu Leif Kobbelt, J.: Structure recovery via hybrid variational surface approximation. Computer Graphics Forum 24(3), 277–284 (2005)
10. Julius, D., Kraevoy, V., Sheffer, A.: D-charts: Quasi-developable mesh segmentation. Computer Graphics Forum 24(3), 581–590 (2005)
11. Chiosa, I., Kolb, A.: Variational multilevel mesh clustering. In: International Conference on Shape Modeling and Applications, pp. 197–204 (2008)
12. Lloyd, S.: Least squares quantization in PCM. IEEE Transactions on Information Theory 28(2), 129–137 (2003)
13. Brox, T., Rosenhahn, B., Cremers, D., Seidel, H.P.: Nonparametric density estimation with adaptive, anisotropic kernels for human motion tracking. In: Elgammal, A., Rosenhahn, B., Klette, R. (eds.) Human Motion 2007. LNCS, vol. 4814, pp. 152–165. Springer, Heidelberg (2007)

14. Abramson, I.S.: On bandwidth variation in kernel estimates a square root law. The Annals of Statistics 10(4), 1217–1223 (1982)
15. Cheng, H., Jiang, X., Sun, Y., Wang, J.: Color image segmentation: advances and prospects. Pattern Recognition 34(12), 2259–2281 (2001)
16. Chen, X., Golovinskiy, A., Funkhouser, T.: A benchmark for 3d mesh segmentation. ACM Trans. Graph. 28, 73:1–73:12 (2009)

Shape Description by Bending Invariant Moments

Paul L. Rosin

School of Computer Science & Informatics, Cardiff University, Cardiff, UK
Paul.Rosin@cs.cf.ac.uk

Abstract. A simple scheme is presented for modifying geometric moments to use geodesic distances. This provides a set of global shape descriptors that are invariant to bending as well as rotation, translation and scale.

Keywords: moments, transformation invariance, bending, articulation.

1 Introduction

In the literature there is a huge range of techniques for shape description, not only for computer vision [14,20,23], but also in particle science [22], medical imaging [3], geography [4], art [1], etc. Shape descriptors can be categorised into global or local methods. Global methods have the advantage of simplicity, and are efficient to both store and match. Local methods provide a richer representation, and so are sometimes more capable of performing object discrimination. However, computing the descriptor tends to require some parameters (e.g. a scale parameter), and matching such descriptors can be computationally expensive.

A desirable property for shape descriptors is that they are invariant to certain transformations of the shape in the image. Invariance to translation, rotation, scale, and possibly skew have become standard practice. However, for shapes that undergo articulation or bending deformation, although there exist methods for matching local descriptors [5,11,15], there is little work on invariant global descriptors [19]. In computer vision textbooks [21] the only instance given is the Euler number, which is generally not sufficiently discriminative.

Moments are widely used as shape descriptors [9,17,18]. Central moments provide translation invariance, and further moment invariants were developed to provide additional invariance to rotation and scale [13] and skew [8]. The contribution of this paper is to describe a simple scheme to enable shapes to be described by geometric moments that are invariant to bending in addition to translation, rotation and scale.

2 Bending Invariant Moments

Our starting point is radial or rotational moments that are generally defined in terms of polar coordinates (r, θ) over the unit disk as

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 253–260, 2011.

$$R_{pq} = \int_0^{2\pi} \int_0^1 r^p e^{iq\theta} f(r\cos\theta, r\sin\theta) r dr d\theta.$$

Since the polar angles of points will vary under bending, we discount these values and just use distances over the 2D image plane. Given a discretised two dimensional shape S, the bending invariant moments are

$$A_p = \sum_{\mathbf{x}_i \in S} d_g(\mathbf{x}_i, \mathbf{x}_c)^p$$

where $\mathbf{x}_c$ is the centre of S. To provide invariance to bending, the Euclidean radial distances have been replaced by geodesic distances. The geodesic distance $d_g(\mathbf{x}_1, \mathbf{x}_2)$ between two points $\mathbf{x}_1$ and $\mathbf{x}_2$ in S is the length of the shortest path between $\mathbf{x}_1$ and $\mathbf{x}_2$ such that the path lies within the S. This path will consist of linear segments and possibly sections of the boundary of S.

The centre also needs to be chosen such that it is invariant to bending. For the Euclidean metric, the centroid is the point minimising the sum of squared distances, while the geometric median is the point that minimises the sum of distances (for which there is no direct closed-form expression). In a similar manner, we choose

$$\mathbf{x}_c = \arg\min_{\mathbf{y} \in S} \sum_{\mathbf{x}_i \in S} d_g(\mathbf{x}_i, \mathbf{y}).$$

Since geodesic distances have been used, the above guarantees that A_p is invariant to translation, rotation and bending. In addition, a normalisation is applied to provide invariance to scale:

$$\alpha_p = \frac{A_p}{A_0^{\frac{p+2}{2}}}$$

where A_0 is the area of S.

Standard geometric image moments of a 2D shape are defined as

$$m_{pq} = \sum_{(x_i, y_i) \in S} x_i^p y_i^q$$

where the distances from the origin are measured along the Cartesian axes. We modify the moments A_p to loosely follow the concept of geometric moments, using something like a set of curvilinear axes. The first axis is the geodesic path to the centre $\mathbf{x}_c$ (used for A_p). The second axis at $\mathbf{x}_i$ is determined locally as the shortest geodesic path from $\mathbf{x}_i$ to the boundary of S. Thus, the second bending and scale invariant moment is

$$\beta pq = \frac{B_{pq}}{B_0^{1+\frac{p+q}{2}}}$$

where

$$B_{pq} = \sum_{\mathbf{x}_i \in S} d_g(\mathbf{x}_i, \mathbf{x}_c)^p d_g(\mathbf{x}_i, E)^q$$

and E is the exterior of S. This provides a more general version of α_p since $\beta_{p0} = \alpha_p$.

To ensure that all points in S are reachable from the centre by geodesic paths we require that S is a single connected component (although holes are allowed). The geodesic centre is not unique – a simple counterexample is given by an annulus. However, this has not proven to be a problem in our experiments.

The only other global bending invariant descriptors of shapes in the computer vision literature that we are aware of is the recent work by Rabin *et al.* [19]. They also use geodesic distances, but they are computed between pairs of points. To reduce computational complexity they perform uniform furthest point sampling of the points in S, and compute geodesics between these points and the full point set in S. Three quartiles of these distances are taken as global descriptors.

Other related work is by Gorelick *et al.* [12]. Instead of the distance transform or geodesic paths they used the Poisson equation which generates expected distances travelled by particles following a random walk. Gorelick *et al.* also computed shape descriptors using moments, but they were not concerned with invariance to bending. Rather than directly computing moments from the random walk distances they used several features derived from the distances. Since several of these features involved the local orientation of the distance field, and combinations of local orientations, the features (and their moments) were not bending invariant. Moreover, the shape was centred at the (standard) centroid which is bending invariant.

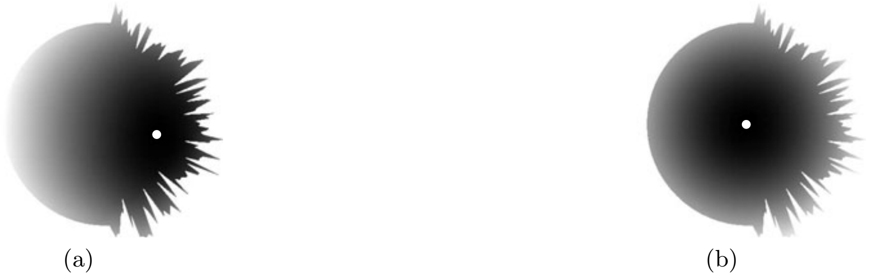
3 Implementation Details

To compute the geodesic distances there are several approaches. Long and Jacobs [15] created a graph whose vertices were sample points. Edges were included in the graph if the straight line between a pair of sample points (corresponding to the vertices) was completely included within the shape. They only used sample points on the boundary, and the cost of both graph construction and determination of the shortest paths between all pairs of points is $O(n^3)$ for n boundary points.

However, in order to reduce the effects of noise, we wish to use geodesic distances between interior points as well as boundary points. This will ensure that the centre will be relatively insensitive to boundary noise. An extreme case is shown in figure 1, in which half the circle has been perturbed. The point minimising the summed geodesic distance to the boundary is significantly affected, unlike the summed geodesic distance to all points in the shape.

We compute geodesic distances using a distance transform capable of operating in non-convex domains. Algorithms for computing the geodesic distance transform using ordered propagation are available with computation complexity linear in the number of pixels [6][1]. Finding the centres of shapes is the most computationally expensive step ($O(n^2)$ for a shape containing n points). Therefore

[1] For simplicity, a geodesic distance transform was used. It performs ordered propagation with a heap based priority queue, and produces approximate geodesic distances.



(a) (b)

Fig. 1. Centres determined by minimising the geodesic distances to (a) the shape boundary, (b) the shape interior. The geodesic distances to the centres are displayed inside the shapes.

a multiresolution strategy is used. The centre is first found at a low resolution after the image has been downsampled so that the shape contains about 5000 pixels[2]. The centre position is then refined at full resolution within a $3F \times 3F$ window, where F is the ratio of high to low resolution. To perform downsampling, each pixel in the low resolution image is set to the foreground value if any of its parent pixels in the high resolution image are foreground. This ensures that the downsampled shape remains a single connected component (even though the topology may change, i.e. holes can be removed).

4 Experiments

We start by showing the effectiveness of the geodesic centre. Despite the shape in figure 2 undergoing articulation and deformation the centre of the shape is captured reliably.

Next, we show classification results for Ling and Jacob's dataset of articulated shapes [15] which consists of 40 shapes made up of 8 classes each containing 5 examples (see figure 3). As a baseline, comparison is made with two methods that are not invariant to bending: the standard Hu moments invariants and Belongie *et al.*'s shape context [2]. In addition, comparison is made with two bending invariant methods: Ling and Jacob's inner-distance (geodesic) extension of shape context [15], and Rabin *et al.*'s work [19] which extracts a 4D description from geodesic quartiles and uses stochastic gradient descent to compute the Wasserstein distance between distributions.

For our classifier we compute the following moments: all A_p of order 1 to 9, all B_{pq} of order 1 to 4 (i.e. 14 moments), and the first 7 Hu moment invariants. For each type of moment we choose the combination of up to 5 moments that

[2] Choosing 5000 pixels for the low resolution version of the shape provides a reasonable compromise between maximising accuracy and computational efficiency. The value was experimentally determined as suitable for all the data tested in this paper.

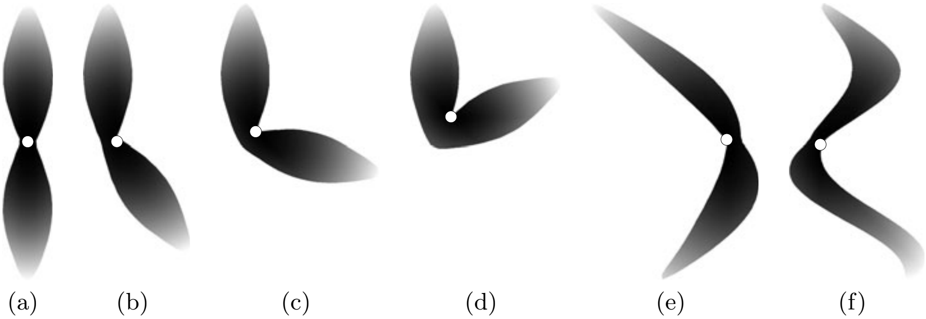


Fig. 2. Despite undergoing articulation and deformation the shape's centre is correctly determined

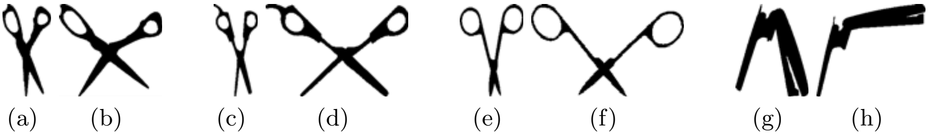


Fig. 3. Examples of pairs of shapes from four classes from Ling and Jacob's dataset of articulated shapes

gives the best classification. Nearest neighbour classification is performed using Mahalanobis distances and leave-one-out cross validation.

The bending moment invariants perform very well, as do the geodesic quartiles, and are only marginally outperformed by the inner distance shape context; see table 1.

Table 1. Classification accuracies for Ling and Jacob's dataset of articulated shapes

Method	**Accuracy (%)**
Hu moment invariants	70
shape context	50
inner distance shape context	100
4D geodesic quartiles	97
α_p	97
β_{pq}	97

Classification performance is evaluated on a second dataset of articulated shapes provided by Gopalan *et al.* [11]. However, since their images contain substantial segmentation errors producing many small holes and components, we have pre-processed the images to extract only the single largest foreground object – examples are shown in figure 4. Gopalan *et al.* report recognition rates

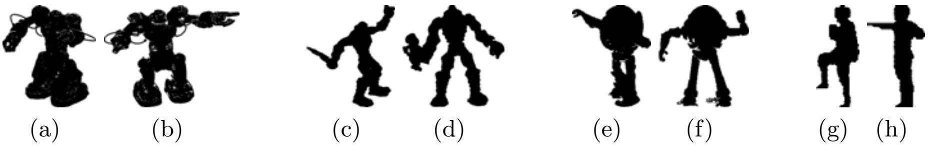


Fig. 4. Examples of pairs of shapes from four classes from Gopalan *et al.*'s dataset of articulated shapes. Only the single largest foreground object from each original image has been retained.

Table 2. Classification accuracies for Gopalan *et al.*'s dataset of articulated shapes

Method	**Accuracy (%)**
Hu moment invariants	72
shape context	82
inner distance shape context	90
α_p	92
β_{pq}	100

Table 3. Bulls-eye test scores for MPEG-7 CE-1 database

Method	**Bulls-eye scores**
Hu moment invariants	46
shape context	76
inner distance shape context	85
ASC & LCDP	96
4D geodesic quartiles	60
α_p	38
β_{pq}	61

of 58% for the inner distance shape context and 80% for their own method, although our experiments (on the cleaned data) using the inner distance shape context gave a much higher accuracy (90%). Again, as table 2 shows, the bending moment invariants perform very well.

Our final evaluation was performed on the MPEG-7 CE-1 database 1400 shapes. As table 3 shows, the invariance to bending does allow β_{pq} to demonstrate clear improvements on the Hu moment descriptors which are only invariant to similarity transformations. However, the bending moment invariants do not perform as well as many other reported methods. The reason is that for some of the object classes there is considerable variation in shape which cannot be captured easily by global descriptors. Note that Rabin *et al.*'s geodesic quartiles give a similar accuracy to β_{pq}. Approaches based on local matching such as shape context can generate excellent accuracies. Currently, the best performance for the MPEG-7 CE-1 database has been achieved by Lin *et al.*'s aspect shape context (ASC) [16] which was combined with locally constrained diffusion process (LCDP) to achieve a bulls-eye score of 96%. However, such methods are computationally expensive.

5 Conclusions

This paper describes an approach to generate global shape descriptors that are invariant to bending, rotation, translation and scale. The shape centre is estimated as the point that minimises the summed geodesic distances to other points, and then moments are computed on the distances from the centre. The benefit of such global shape descriptors is that they tend to be much more efficient to match compared to local shape descriptors. For instance, matching by shape context requires dynamic programming, while both [19] and [5] use iterative matching schemes.

Several approaches in the literature could be adapted to the proposed framework. For instance, for describing 3D shapes Gal *et al.* [10] use a histogram of centricity, which is the average of the geodesic distance from a mesh vertex to all other vertices. The proposed moment descriptors could also be generated from centricity values instead of distances from the centre. However, since computing pairwise distances is computationally expensive this would have the disadvantage of requiring all the pairwise distances to be computed at full resolution.

An alternative to directly using the geodesic distances to modify the moments, is to use the geodesic distances along with multidimensional scaling to construct a bending invariant form of the shape [7,15]. Moments subsequently computed would therefore be bending invariant. Future work will investigate the effectiveness of this approach.

Finally, extending the proposed method to 3D shapes would be straightforward. The geodesic distance transform using ordered propagation can be readily and efficiently applied in higher dimensions, with run-time proportional to the number of elements (e.g. voxels).

References

1. Arnheim, R.: Art and Visual Perception: A Psychology of the Creative Eye. University of California Press, Berkeley (1974)
2. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. on Patt. Anal. and Mach. Intell. 24(4), 509–522 (2002)
3. Bookstein, F.L.: Shape and the information in medical images: A decade of the morphometric synthesis. Computer Vision and Image Understanding 66(2), 97–118 (1997)
4. Boyce, R.B., Clark, W.A.V.: The concept of shape in geography. Geographical Review 54, 561–572 (1964)
5. Bronstein, A.M., Bronstein, M.M., Bruckstein, A.M., Kimmel, R.: Analysis of two-dimensional non-rigid shapes. Int. J. of Computer Vision 78(1), 67–88 (2008)
6. Cárdenes, R., Alberola-López, C., Ruiz-Alzola, J.: Fast and accurate geodesic distance transform by ordered propagation. Image Vision Comput. 28(3), 307–316 (2010)

7. Elad, A., Kimmel, R.: On bending invariant signatures for surfaces. IEEE Trans. on Patt. Anal. and Mach. Intell. 25(10), 1285–1295 (2003)
8. Flusser, J., Suk, T.: Pattern recognition by affine moment invariants. Pattern Recognition 26, 167–174 (1993)
9. Flusser, J., Zitova, B., Suk, T.: Moments and Moment Invariants in Pattern Recognition. Wiley Publishing, Chichester (2009)
10. Gal, R., Shamir, A., Cohen-Or, D.: Pose-oblivious shape signature. IEEE Trans. Vis. Comput. Graph. 13(2), 261–271 (2007)
11. Gopalan, R., Turaga, P.K., Chellappa, R.: Articulation-invariant representation of non-planar shapes. In: Europ. Conf. Computer Vision, vol. 3, pp. 286–299 (2010)
12. Gorelick, L., Galun, M., Sharon, E., Basri, R., Brandt, A.: Shape representation and classification using the poisson equation. IEEE Trans. on Patt. Anal. and Mach. Intell. 28(12), 1991–2005 (2006)
13. Hu, M.: Visual pattern recognition by moment invariants. IRE Trans. Inf. Theory 8(2), 179–187 (1962)
14. Kindratenko, V.V.: On using functions to describe the shape. J. Math. Imaging Vis. 18(3), 225–245 (2003)
15. Ling, H., Jacobs, D.W.: Shape classification using the inner-distance. IEEE Trans. on Patt. Anal. and Mach. Intell. 29(2), 286–299 (2007)
16. Ling, H., Yang, X., Latecki, L.J.: Balancing deformability and discriminability for shape matching. In: Europ. Conf. Computer Vision, vol. 3, pp. 411–424 (2010)
17. Mukundan, R., Ramakrishnan, K.R.: Moment Functions in Image Analysis – Theory and Applications. World Scientific, Singapore (1998)
18. Prokop, R.J., Reeves, A.P.: A survey of moment-based techniques for unoccluded object representation and recognition. CVGIP: Graphical Models and Image Processing 54(5), 438–460 (1992)
19. Rabin, J., Peyré, G., Cohen, L.D.: Geodesic shape retrieval via optimal mass transport. In: Europ. Conf. Computer Vision, vol. 5, pp. 771–784 (2010)
20. Rosin, P.L.: Computing global shape measures. In: Chen, C.H., Wang, P.S.-P. (eds.) Handbook of Pattern Recognition and Computer Vision, 3rd edn., pp. 177–196. World Scientific, Singapore (2005)
21. Sonka, M., Hlavac, V., Boyle, R.: Image Processing, Analysis, and Machine Vision. Thomson-Engineering (2007)
22. Taylor, M.A.: Quantitative measures for shape and size of particles. Powder Technology 124(1-2), 94–100 (2002)
23. Zhang, D., Lu, G.: Review of shape representation and description techniques. Pattern Recognition 37(1), 1–19 (2004)

Fast Shape Re-ranking with Neighborhood Induced Similarity Measure

Chunyuan Li[1], Changxin Gao[1], Sirui Xing[1], and Abdessamad Ben Hamza[2]

[1] Huazhong University of Science and Technology, China
[2] Concordia Institute for Information Systems Engineering,
Concordia University, Montréal, QC, Canada

Abstract. In this paper, we address the shape retrieval problem by casting it into the task of identifying "authority" nodes in an inferred similarity graph and also by re-ranking the shapes. The main idea is that the average similarity between a node and its neighboring nodes takes into account the local distribution and therefore helps modify the neighborhood edge weight, which guides the re-ranking. The proposed approach is evaluated on both 2D and 3D shape datasets, and the experimental results show that the proposed neighborhood induced similarity measure significantly improves the shape retrieval performance.

Keywords: Shape retrieval, graph theory, similarity, re-ranking.

1 Introduction

Searching shapes more accurately and faster is one of the most important goals in computer vision. In recent years, several approaches have been proposed to optimize shape retrieval systems, from designing smart descriptors [1–4], to exploring suitable similarity measure-based methods [5]. However, almost all these systems suffer from the following phenomenon: when a user submits a query, some shapes in the database are returned relatively often, while some are returned only given certain special queries. To tackle this problem, we propose the Neighborhood Induced Similarity (NIS), which updates the original similarity based on the neighborhood of a shape before the final ranking. Traditional shape retrieval systems compute the pair-wise similarity among shapes, from which a global ordering can be derived. By separating ranking from similarity measurements, one can leverage ranking algorithms to generate a global ordering. Just like existing re-ranking algorithms for web page [6], image [7] and video [8] retrieval, our proposed method also takes the similarity/dissimilarity/distance matrix as the input, and outputs an optimized similarity matrix for the final ranking. However, the difference in our setting is that we aim at eliminating the undesired phenomenon stated above. We present an approach from the perspective of a graph representation, where shapes are represented as graph nodes and the graph edges encode the similarity between these shapes. In this way, searching shapes is formulated as label propagation on a graph. Therefore, the edge value is very critical, since it guides the propagation behavior. In our method,

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 261–268, 2011.

we consider all the nodes on the graph as a whole, and update the edge value by the average similarity of nodes, resulting in an improved retrieval accuracy.

Shape retrieval techniques may be broadly classified into two main categories: traditional matching/retrieval methods and similarity learning methods [5]. Belongie *et al.* [1] introduced shape contest, which is a 2D histogram representation of a shape. Ling and Jacobs [2] proposed the inner distance which modifies shape contexts by considering the geodesic distance between contour points instead of the Euclidean distance, and thus significantly improves the retrieval and classification of articulated shapes. Wei *et al.* [9] extracted Zernike features to describe trademark shapes. Trademark images are very complex 2D shapes, and obtaining a high performance of trademark retrieval is of paramount importance to the industry. On the other hand, the Light Field Descriptor (LFD) [10] has been reported in the literature as one of the most efficient techniques for retrieving 3D rigid models [11]. Our method is, however, general and not limited to any particular similarity measure or representation.

Our work is partly motivated by Bai *et al.'s* work [5], which adopts a graph-based transductive learning algorithm to improve the shape retrieval results. The key idea of this distance learning algorithm is to replace the original shape distance with a distance induced by geodesic paths on a manifold of known shapes. In other words, each shape is considered in the context of other shapes in its class, and the class need not be known. Instead of Graph Transduction, we propose in this paper a neighborhood induced similarity to improve the shape retrieval results. There has been a significant body of work on similarity measures-based methods. Cheng *et al.* [12] proposed the sparsity induced similarity measure to improve the label propagation performance [13]. Jegou *et al.* [14] proposed the Contextual Dissimilarity Measure (CDM) to improve the image search accuracy through improving the symmetry of the k-neighborhood relationship. Our work reassigns edge weight in a fully connected graph by the average neighborhood weight, which is in a similar spirit as CDM. More related work on similarity measure-based methods can be found in [15]. The main advantages of our proposed approach may be summarized as follows:

- It eliminates the abnormal frequency phenomenon for shape retrieval with a graph representation.
- It is universal to arbitrary 2D and 3D shapes.

The rest of the paper is organized as follows. The problem formulation is stated in Section 2. The proposed neighborhood induced similarity approach is described in Section 3. The experimental results using the proposed algorithm are provided in Section 4. Finally, we conclude in Section 4 and point out future work directions.

2 Problem Formulation

In traditional shape retrieval systems, a user usually employs a distance function to compute the pair-wise similarity between two shape features, and assumes that the more similar two shapes are, the smaller their difference is. For a given query, these systems rank the shapes in the dataset as a list according to the values of the pair-wise similarity, and present to the user several top rankings in the returned list. To measure the number of returned times of a shape in a given dataset, we first introduce the concept of *appearing frequency* of a shape. Suppose there are N shapes in a dataset, and let us consider the top t ranking shapes $R_t(n)$ in the returned list $L(n)$ of a query shape $Q_n, 1 \leq n \leq N$. Obviously, the cardinality $|R_t(n)| = t$ of the set $R_t(n)$ is constant within the t-highest ranking framework. The appearing frequency of Q_n is then defined as follows:

$$f(n) = \sum_{i=1}^{N} \sum_{j \in R_t(n)} \delta_{i,j} \tag{1}$$

where

$$\delta_{i,j} = \begin{cases} 1, & \text{if } Q_n \text{ is the } j\text{-th returned shape of the query } Q_i \\ 0, & \text{otherwise.} \end{cases}$$

We can observe that some shapes have high frequency rate, while others are returned only when submitting specific queries. These shapes are referred to as *over-returned shapes* and *never-returned shapes* respectively, which are defined for a given neighborhood size. Both of them are considered abnormal shapes or 'bad shapes' in a shape retrieval system. Our goal is to reduce the number of these abnormal shapes. In other words, we hope that the frequency rates of each shape in the dataset would be the same constant which is relative to $|R_t(n)|$.

3 Proposed Neighborhood Induced Similarity Approach

Let $D = (d_{ij})$ be a distance matrix computed by a shape function. We formulate the shape retrieval problem as a form of propagation on a graph, where a node's label propagates to the neighboring nodes according to their proximity. In this process, we fix the label on the query shape. Thus, the query shape acts like a source that pushes out the label to other shapes. Intuitively, we want shapes that are similar to have the same label. We create a graph $G = (\mathcal{V}, \mathcal{E})$ where the node set $\mathcal{V}$ represents all the shapes in the dataset, both query and the others. The elements of the edge set $\mathcal{E}$ represent the similarity between nodes (shapes). We propagate the labels through the edges. Larger edge weights allow labels to travel through more easily. The propagation process stops when a user-specified number of nodes are labeled. Likewise, the frequency of a node $\boldsymbol{v}_i \in \mathcal{V}$ is defined as the sum of the labeled times after each node in the graph propagates its label to its neighborhood. Interestingly, we find that the frequency of a node is equal to the degree of the node if we cut off the edges that no label travels. A subgraph of the newly obtained graph G_0 includes a dense graph G_{dense} and a sparse graph G_{sparse}. Obviously, G_{dense} consists of nodes with high frequency (or degree) V_{high}, and G_{sparse} consists of nodes with low frequency V_{low}. Viewed in this fashion, our target is to obtain a well-distributed graph.

Now assume that the graph is fully connected with the weights computed by a Gaussian kernel as follows:

$$w(\boldsymbol{v}_i, \boldsymbol{v}_j) = \exp\left(-\frac{d_{ij}}{\alpha^2}\right), \tag{2}$$

where α is a bandwidth parameter that is usually determined empirically. In the sequel, we set the parameter α to 100.

The k-nearest neighbors of a given node $\boldsymbol{v}_i$ are the nodes $NN_k(i)$, in which the nodes $\boldsymbol{v}$ and $\boldsymbol{v}_i$ are connected by an edge, if the edge weight between $\boldsymbol{v}$ and $\boldsymbol{v}_i$ is among the k-th largest from $\boldsymbol{v}_i$ to other nodes, i.e.

$$NN_k(i) = \{\boldsymbol{v} : \max_k w(\boldsymbol{v}_i, \boldsymbol{v})\}. \tag{3}$$

The above-mentioned problem of frequency rate suggests a solution which reassigns weight. Intuitively, we would like the k-neighborhoods to have similar weights in order to eliminate ‘bad shapes’.

Let us consider the neighborhood of a given node defined by its $|NN_k(i)|$ nearest neighbors. The value k is a compromise between computation cost and quality of retrieval result. The larger the value of k, the more expensive the computation is. In general, k needs to be greater than 20 to prevent the system from being over-constrained due to possible noise in the original measure.

We define the neighborhood weight or similarity $s(i)$ as the mean weight of a given node $\boldsymbol{v}_i$ to the nodes of its neighborhood:

$$s(i) = \frac{1}{k} \sum_{\boldsymbol{x} \in NN_k(i)} w(\boldsymbol{v}_i, \boldsymbol{x}) \tag{4}$$

and it is computed for each node. Subsequently, we define a new weight between two nodes as follows:

$$w^\star(\boldsymbol{v}_i, \boldsymbol{u}_k) = w(\boldsymbol{v}_i, \boldsymbol{u}_k) \frac{\bar{s}}{(s(i)s(j))^{1/2}}, \tag{5}$$

where $\bar{s}$ is the geometric mean neighborhood similarity obtained by

$$\bar{s} = \prod_i s(i)^{1/n}. \tag{6}$$

Thus, we reassign the graph weight and propagate the query label according to the new weight. Note that the terms $\bar{s}$ and $s(i)$ do not impact the nearest neighbors of a given node.

4 Experimental Results

In the first experiment, we test the performance of the proposed NIS approach for improving the retrieval results on 2D Wei’s Trademark Image dataset which consists of 1003 images in 14 classes [9]. These include Apple, Fan, Abstract

Circle1, Abstract Circle2, Encircled Cross, Abstract Saddle, Abstract Saddle, Abstract Sign, Triangle, Bat, Peacock, Abstract Flowers, Rabbit, Snow Flake and Miscellaneous. Fig. 1 displays sample images from this Trademark Image dataset. Traditional descriptor-based methods applied on these complex shapes still maintain a modest performance. However, the industry is in urgent need for a satisfactory retrieval performance because it will save human consumption of comparing trademark shapes one by one to avoid reduplication. In this experiment, the neighborhood size is set to 85.

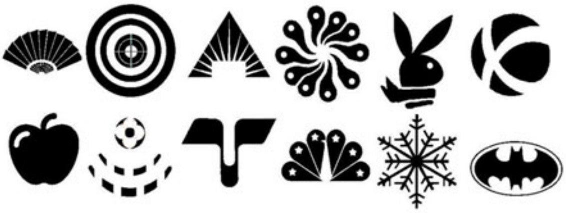


Fig. 1. Sample images from the trademark shape dataset

In our comparative analysis, we used the Precision/Recall curve to measure the retrieval performance. Ideally, this curve should be a horizontal line at unit precision. For each query image, we use the first 108 return trademark images with descending similarity rankings (i.e. ascending Euclidean distance ranking), dividing them into 9 groups accordingly. However, in order to obtain a more objective picture of the performance, we plot the average performance of 20 query images of the same class. We show the retrieval result in Fig. 2, where Zernike features [3] are first extracted to calculate the original distance matrix, and then the proposed algorithm is used to obtain the accuracy-improved matrix. Again, the overall retrieval performance is improved by NIS.

We also tested the performance of the proposed matching algorithm on the McGill Shape Benchmark [16]. This publicly available benchmark database provides a 3D shape repository, which contains 255 objects that are divided into ten categories, namely, 'Ants', 'Crabs', 'Spectacles', 'Hands', 'Humans', 'Octopuses', 'Pliers', 'Snakes', 'Spiders', and 'Teddy Bears'. Sample models from this database are shown in Fig. 3. In this experiment, we set k to 30.

The evaluation of the 3D retrieval results is based on the following quantification measures. These measures range from 0% to 100%, and higher values indicate better performance.

A. Nearest Neighbor (NN) The percentage of queries where the closest match belongs to the query's class.

B. First Tier (FT): the recall for the $(\kappa - 1)$ closest matches, where κ is the cardinality of the query's class.

C. Second Tier (ST): the recall for the $2(\kappa - 1)$ closest matches.

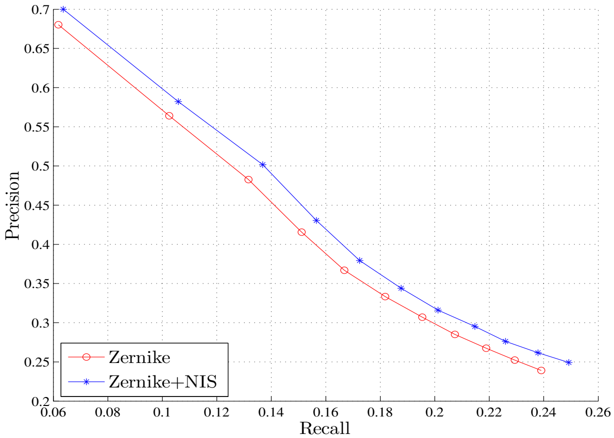


Fig. 2. NIS improves the overall retrieval performance for the trademark image dataset. Blue line is with NIS and red line is without NIS.

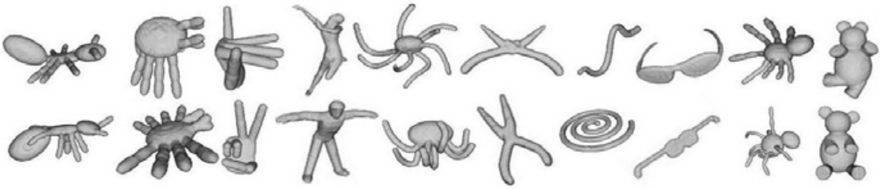


Fig. 3. Sample shapes from McGill Articulated Shape Database. Only two shapes for each of the 10 classes are shown.

D. Discounted Cumulative Gain: a statistic that emphasizes on retrieving highly relevant shapes. Correct results near the front of the retrieval list are weighted more heavily than correct results near the end, under the assumption that a user is most interested in the first results.

We compare our method with the 3D Light Field Distribution (LFD) method. The distance matrix calculated via LFD is chosen as the input of NIS, and the neighborhood size is set to 48. The corresponding scores of each method for each class of the database as well as the overall scores for the complete database are shown in Table 1. We render numbers in bold if our method is superior or equivalent to LFD. Obviously, in most cases, NIS has a positive effect on 3D shapes re-ranking. Though not all the entries with NIS outperform LFD in the comparative results, it is understandable that NIS works in a given statistical range. For example, 'Hands', the worst class according to the result, will still lead to effective shape re-ranking based upon the First Tier measure.

The time to compute the neighborhood distance is $\mathcal{O}(kN)$, to compute geometric mean is $\mathcal{O}(N)$, and to update the weight is $\mathcal{O}(N)$. There is an additional cost for ranking the scores at the end, which is $\mathcal{O}(k \log N)$. Thus, the total

Table 1. Quantitative measure scores of the retrieval methods

# Queries	**Method**	**NN (%)**	**FT (%)**	**ST (%)**	**DCG (%)**
Overall Database	Ours	84.16	**46.26**	**62.18**	**83.72**
	LFD	84.61	44.69	59.28	82.74
Ants	Ours	90	53.6	**77.78**	88.42
	LFD	93.33	54.56	76.11	89.33
Crabs	Ours	**93.33**	**50.89**	**65.33**	**86.33**
	LFD	93.33	45	60.11	84.08
Spectacles	Ours	76	**51.52**	**66.08**	**88.45**
	LFD	100	50.56	65.60	88.34
Hands	Ours	80	**28.25**	40.50	75.36
	LFD	90	28	42.25	75.44
Humans	Ours	**79.31**	**40.79**	**54.94**	**81.26**
	LFD	79.31	39.35	53.03	80.69
Octopuses	Ours	**60**	**26.88**	**41.93**	**72.30**
	LFD	48	24	35.36	68.47
Pliers	Ours	**100**	**75.50**	**87.25**	**97.58**
	LFD	100	75.25	87.25	97.47
Snakes	Ours	**76**	**26.14**	**33.92**	**71.68**
	LFD	68	20.64	25.28	66.53
Spiders	Ours	70.97	41.94	**65.45**	**81.18**
	LFD	74.12	42.77	65.35	82.35
Teddy Bears	Ours	**100**	66.50	**86.50**	**95.64**
	LFD	100	66.75	82.50	94.59

complexity of our method amounts to $\mathcal{O}(k \log N + (k+2)N)$. Since $k \ll N$, the overall time complexity of our algorithm is bounded by $\mathcal{O}(N^2)$. It is worth pointing out that the complexity of our approach is at least one order smaller than the complexity $\mathcal{O}(TN^3)$ of the graph transduction algorithm, where T is the number of iterations.

5 Conclusions and Future Work

In this paper, we proposed a novel similarity measure for eliminating abnormal shapes in shape retrieval systems. The proposed NIS measure takes into account the hidden local structure by using the average neighborhood similarity in a graph representation. We tested the proposed similarity measure on 2D and 3D shape datasets. The experimental results demonstrated the efficiency of the proposed method both in 2D and 3D, even on shapes with large variations. Future research directions include further exploration of the frequency problem on shape classification as well as clustering. We also plan to combine sparse representation with the proposed method in order to achieve much better retrieval results.

References

1. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Analysis and Machine Intelligence 24(4), 509–522 (2002)
2. Ling, H., Jacobs, D.: Shape classification using the inner-distance. IEEE Trans. Pattern Analysis and Machine Intelligence 29(2), 286–299 (2007)
3. Li, S., Lee, M.-C., Pun, C.-M.: Complex Zernike moments features for shape-based image retrieval. IEEE Trans. Systems, Man, and Cybernetics 39(1), 227–237 (2009)
4. Sebastian, T.B., Klein, P.N., Kimia, B.B.: Recognition of shapes by editing their shock graphs. IEEE Trans. Pattern Analysis and Machine Intelligence 26(5), 550–571 (2004)
5. Bai, X., Yang, X., Latecki, L.J., Liu, W., Tu, Z.: Learning context sensitive shape similarity by graph transduction. IEEE Trans. Pattern Analysis and Machine Intelligence 32(5), 861–874 (2010)
6. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. In: Proc. Int. Conf. World Wide Web 7, vol. 30(1-7), pp. 107–117 (1998)
7. He, X., Ma, W.-Y., Zhang, H.: Imagerank: spectral techniques for structural analysis of image database. In: Proc. IEEE Int. Conf. Multimedia and Expo, vol. 1, pp. 25–28 (2002)
8. Latecki, L., Lakamper, R., Eckhardt, U.: Shape descriptors for non-rigid shapes with a single closed contour. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 424–429 (2000)
9. Wei, C.-H., Li, Y., Chau, W.-Y., Li, C.-T.: Trademark image retrieval using synthetic features for describing global shape and interior structure. Pattern Recognition 42(3), 386–394 (2009)
10. Chen, D.-Y., Tian, X.-P., Shen, Y.-T., Ouhyoung, M.: On visual similarity based 3D model retrieval. Computer Graphics Forum 22(3), 223–232 (2003)
11. Siddiqi, K., Zhang, J., Macrini, D., Shokoufandeh, A., Bouix, S., Dickinson, S.: Retrieving articulated 3-D models using medial surfaces. Machine Vision and Applications 19(4), 261–275 (2008)
12. Cheng, H., Liu, Z., Yang, J.: Sparsity induced similarity measure for label propagation. In: Proc. IEEE Int. Conf. Computer Vision, Kyoto, Japan, pp. 317–324 (2009)
13. Xiaojin, Z.: Semi-supervised learning with Graphs, PhD thesis, CMU (2005)
14. Jegou, H., Schmid, C., Harzallah, H., Verbeek, J.: Accurate image search using the contextual dissimilarity measure. IEEE Trans. Pattern Analysis and Machine Intelligence 32(1), 2–11 (2010)
15. Yu, J., Amores, J., Sebe, N., Radeva, P., Tian, Q.: Distance learning for similarity estimation. IEEE Trans. Pattern Analysis and Machine Intelligence 30(3), 451–462 (2008)
16. http://www.cim.McGill.ca/shape/benchmark

Dynamic Radial Contour Extraction by Splitting Homogeneous Areas

Christopher Malon and Eric Cosatto

NEC Laboratories America
lastname@nec-labs.com

Abstract. We introduce a dynamic programming based algorithm to extract a radial contour around an input point. Unlike many approaches, it encloses a region using feature homogeneity, without relying on edge maps. The algorithm operates in linear time in the number of pixels to be analyzed. Multiple initializations are unnecessary, and no fixed smoothness/local–optimality tradeoff needs to be tuned. We show that this method is beneficial in extracting nuclei from color micrographs of hematoxylin and eosin stained biopsy slides.

Keywords: Contour extraction, dynamic programming, histological images, segmentation, digital pathology.

1 Introduction

Contour extraction quality can critically affect image features used for pattern recognition and classification. Histological imaging, in particular, may require fast, sensitive feature measurements of thousands of cells. As the goal may be to detect irregular cells, outlying measurements cannot simply be smoothed over.

In the nascent practice of digital pathology [1], this problem manifests itself in the detection and grading of breast cancer based on scanned micrographs of hematoxylin and eosin stained biopsy tissues. Variation in the size, shape, and texture of nuclei determine the degree of *pleomorphism*, which is a component of the Bloom–Richardson grade [2] and an indicator of malignancy itself. These attributes rely on accurate nuclear contours, yet the general purpose dyes often fail to produce crisp edges at nuclear boundaries.

For this reason, we aim to develop a contour extraction system that can fit the nuclei with less sensitivity to edge sharpness. We propose a very general framework that performs global *homogeneity splitting*. The framework replaces the traditional edge signal with differences in texture over an area, measured all the way back to the center of a nucleus. It chooses a globally optimal boundary using a dynamic programming algorithm. Overall computation time is linear in the number of pixels analyzed.

Nuclear extraction has been studied frequently in the cytology and emerging digital pathology literature. In comparison to our proposed approach, edge–based methods typically require multiple filter strengths to describe all contours: [3] uses adaptive thresholding and [4] uses a Watershed algorithm. In liquid cytology,

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 269–277, 2011.

some authors use snakes [5], but some work using dynamic programming search has emerged [6]. Compared to [6], our objective uses texture homogeneity instead of edge fitting, it has lower complexity, and it enforces continuity by limiting the size of a radial jump instead of allowing arbitrary jumps penalized by elasticity cost.

Radial contour extraction also has appeared in the radiology literature. In [7], a typical dynamic programming algorithm [8] is applied in the space of polar coordinates, to extract the radial contour of a ventricule. Their algorithm lacks the loop constraint that we impose, and large radial jumps to close the contour would not be penalitzed. Also, in contrast to our homogeneity objective, the local objective of their algorithm is to maximize the response of an edge filter.

Our radial contour extraction algorithm most closely resembles the multiple backtracking algorithm of Sun and Pallottino [9]. Ours is slightly more general, as we allow the radius to jump up to a constant distance between discretized angles. Although Sun and Pallottino did not prove that their algorithm always succeeded in producing a closed loop, we prove that our algorithm does in Theorem 1. Later work of Appleton and Sun [10] gives a different solution to radial contour extraction that yields the true minimum cost path at more time expense in the worst case but just as quickly as dynamic programming in the average case.

Segmentation techniques, such as Seeded Region Growing [11], iteratively enlarge regions by comparing pixels to be enclosed to the average of a feature already measured. Our dynamic programming approach solves for a global optimum, so it will not be perplexed by aberrant pixels. The minimization of a second–order statistic over the enclosed region establishes another remarkable difference.

2 Homogeneity Splitting

The first novelty of our approach is the objective of the contour extraction. We do not apply an edge filter to the original signal.

Rather, we take an arbitrary signal $T(x, y)$ presumed to be homogeneous inside the desired boundary (here we simply use pixel luminance, but more complicated texture features could be used interchangeably). Homogeneity of T around the point $(0, 0)$ inside a boundary parameterized in polar coordinates by $\rho(\theta)$ is defined by variance in T:

$$H = \int_{\theta=0}^{2\pi} \sigma_\theta(\rho(\theta)) d\theta \tag{1}$$

$$\text{where } \sigma_\theta(r) = \int_{r'=0}^{r} (T_{polar}(r', \theta))^2 dr' - \frac{1}{r}\left(\int_{r'=0}^{r} T_{polar}(r', \theta) dr'\right)^2 \tag{2}$$

Our goal is to bound the center (say $(0, 0)$) of the detection by a boundary B, such that the variance H of T inside B is small, and sharply rises if B is extended. This amounts to minimizing a local cost function

$$C_\theta(r) = \sigma_\theta(r) - \sigma_\theta(r + \delta) + 1 \tag{3}$$

for a constant $\delta = lookAhead$ that affects smoothness. Assuming that F is normalized so that $0 \leq F \leq 1$, the “1” term ensures that $C_\theta \geq 0$.

3 Radial Contour Extraction

The dynamic radial contour extraction produces a contour around each input point. The main steps of the algorithm are summarized in Table 1.

Table 1. Dynamic Radial Contour Algorithm

```
Input.
    Real–valued feature map $F(x, y)$, $0 \leq F(x, y) \leq 1$
    Point $(p, q)$
    Integers $minRadius$, $maxRadius$, $numTheta$, $lookAhead$, $maxJump$
Algorithm.
    $G \leftarrow \mathrm{PolarTransform}(F, maxRadius, numTheta, p, q)$
    $C \leftarrow \mathrm{VarianceCost}(G)$
    $B \leftarrow \mathrm{LoopDP}(G, C, minRadius, lookAhead)$
Output. $\mathrm{RectTransform}(B, maxRadius, numTheta, p, q)$
```

First, a polar transform around the input point transforms the feature map F into a map G:

$$G(r, s) = F(p + r \cos\theta, q + r \sin\theta) \tag{4}$$

where

$$\theta = \frac{2\pi s}{numTheta}\,;\ 0 \leq s \leq numTheta\,;\ 1 \leq r \leq maxRadius \tag{5}$$

The contours to be produced by our algorithm may be parameterized by their radii $\rho = r(\theta)$ in terms of the angle (ı.e., they do not cross the same angle multiple times). At each angle θ, we measure homogeneity along the arc from (p, q) out to the point $r(\theta)$ by the standard deviation in G:

$$\sigma_\theta(r) = \sqrt{\frac{\sum_{r'=1}^{r} G(r', s)^2}{r^2} - \left(\frac{\sum_{r'=1}^{r} G(r', s)}{r}\right)^2}$$

Locally (at θ), we wish to draw the boundary at a point where the homogeneity inside the contour is good ($\sigma_\theta(r)$ is small) but the homogeneity would be much worse if the contour were pushed further out ($\sigma_\theta(r + lookAhead)$ is much bigger). Thus, we define the local cost

$$C(r, s) = \sigma_\theta(r) - \sigma_\theta(r + lookAhead) + 1 \tag{6}$$

and the global cost

$$Cost(\rho) = \sum_{s=0}^{numTheta-1} C_\theta(\rho(\frac{2\pi s}{numTheta})). \tag{7}$$

The range of F ensures that the cost will be nonnegative.

The *LoopDP* function solves a relaxation of the following minimization problem: Determine ρ minimizing $Cost(\rho)$, such that $\rho(\theta) \in [minRadius, maxRadius]$, $|\rho(\frac{2\pi(s+1)}{numTheta}) - \rho(\frac{2\pi s}{numTheta})| \leq maxJump$, and $\rho(0) = \rho(2\pi)$. Setting an appropriate $minRadius$ avoids a trivially small enclosure with zero variance in F. LoopDP is implemented by a dynamic programming algorithm [12]. With complexity $O(numTheta \cdot maxRadius \cdot maxJump)$ the typical algorithm gives the cheapest assignments $\rho(\theta(s))$ ending at $G(r_2, numTheta)$, without restriction on the originating node $G(r_1, 0)$. A modification to the typical algorithm rejects solutions where $r_1 \neq r_2$ by imposing an infinite cost at the final step of a path to $(numTheta, r_2 = \rho(2\pi))$ when $r_1 \neq r_2$. Also, at any step where there are equal costs, the algorithm always prefers the smaller–indexed parent.

Theorem 1. *The modified algorithm finds a path with finite cost.*

(For the proof, see section 6.) This path translates to a closed loop in rectangular coordinates.

The overall complexity of our algorithm (the computation of C, followed by the dynamic programming search) is $O(maxRadius \cdot numTheta \cdot maxJump \cdot lookAhead)$ per candidate. As $maxJump$ and $lookAhead$ enforce continuity of the extracted contours, they should be regarded as constant (here we set $maxJump = 1$). If candidates are well–separated (meaning that the neighborhoods around the candidates do not overlap, on average), then the whole procedure may be completed in linear time, relative to the number of pixels in the input image. Our method is faster than typical implementations of the active contour algorithm, which run for at least several hundred and maybe several thousand iterations to produce a single contour (and many initializations are typically used), and consider many initializations.

The given contour extraction algorithm will return a contour around any input point. As a simple method to filter bad candidate detections, we propose setting a threshold on H inside the extracted contour.

4 Experiments

The baseline for our experiments is our previous state–of–the–art active contour, edge–based contour extraction algorithm, described in [13], in which bad extractions are rejected by an SVM. We are interested in the ability of the homogeneity–based dynamic radial contour algorithm to pick up nuclei missed by the baseline algorithm. Furthermore, we are interested in the significance of these detections for cancer diagnosis.

The present algorithm extracts a contour to enclose each *nuclear site*. The nuclear sites are provided by another algorithm as input to this contour extractor. Here we use a Difference of Gaussian (DoG) method. Three DoG filters are used, to extract small, medium, and large nuclei. As in [13], the sites for the active contour method may also be selected as peaks of a Hough transform [14].

As for breast cancer, gastric cancer diagnosis weighs heavily on the shape and distribution of cell nuclei [15]. Here we study a set of 588 regions of interest (ROI) from 453 distinct slides of biopsy tissue of patients suspected of gastric cancer. The tissues were stained by hematoxylin and eosin, and a computed selected microscopic regions of interest, measuring 233 microns by 233 microns, or 1024 pixels by 1024 pixels at 400X objective magnification. At this magnification, a Japanese pathologist diagnosed whether cancer appeared in each individual ROI, and the computer was challenged to make the same decisions.

The difference in extraction methods may be seen in Figure 1. If the active contour method is used together with a tight SVM threshold on nuclear detections, as intended in [13], very few nuclei can be segmented (image (b)). If a looser threshold is used, so that the number of outputs matches the number of outputs of the Homogeneity Splitting method overall, then a substantial difference in quality may be seen (images (c) and (d)).

Quantitatively, we show the diagnostic importance of the different extractions by attempting to perform cancer diagnosis using features derived from the nuclei extracted by the present method (HS) and the method of [13] (AC). The results reflect 3–fold cross validation over the 588 ROI set. Correlations among ROI from the same slides are substantial, so ROI from each slide were either taken into or omitted from each set of the partition as a unit.

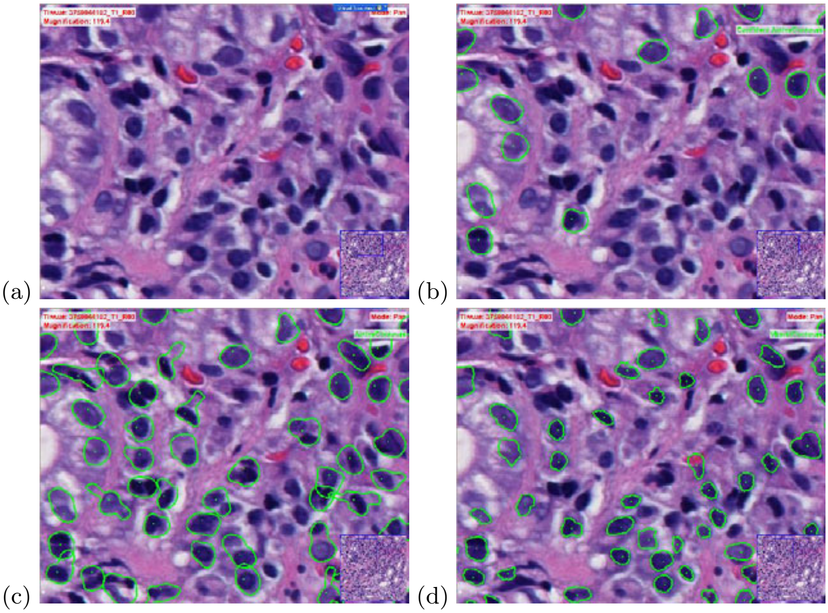


Fig. 1. Comparison of contour extraction algorithms. (a) Original image; (b) Active Contour detections (tight threshold); (c) Active Contour detections (loose threshold); (d) the present algorithm (loose threshold).

Our results are summarized in the first part of Table 2. For fair comparison, each contour method should be compared at a threshold producing the same number of detections. Hence Table 2 has four entries. The method of thresholding the detections differs between contour methods (the AC method uses a support vector machine on fourteen features; the HS method uses a threshold on luminance homogeneity). At a "tight" threshold, each extractor is run on large nuclei detections; at the loose threshold, all three DoG filters are used. At the "tight" threshold, about 1,000 large nuclei are extracted by each method (50,000 nuclei of all sizes). At the "loose" threshold, about 270,000 nuclei (of all sizes) are extracted. The loose threshold was not carefully tuned, but rather chosen to exclude some obvious misdetections. The recognition rate reported in the table is "balanced recognition": one–half the recognition rate on ROI labeled as cancerous by the pathologist, plus one–half the recognition rate on ROI labeled as non-cancerous.

Table 2. Classification results with different contour techniques

Problem		Contour	Threshold	Recognition Rate
ROI	1.	AC	Tight (large nuclei)	67.2%
ROI	2.	HS	Tight (large nuclei)	54.0%
ROI	3.	AC	Loose (all nuclei)	69.4%
ROI	4.	HS	Loose (all nuclei)	**70.8%**
Gland	1.	AC	Loose (all nuclei)	62.0%
Gland	2.	HS	Loose (all nuclei)	**65.7%**

At the tight threshold, the AC method is clearly better at producing a few significant high–confidence contours. At the loose threshold, the HS method performs better. It produces more reliable nuclear contours when it is not acceptable to reject many nuclei.

Gland classification. The problem of gland nuclear analysis makes it even clearer that the HS method produces more reliable nuclear contours when it is not acceptable to reject many nuclei. Here, a separate algorithm [16] automatically extracts binary masks representing epithelia of glands. There should be only one mask per gland, and the algorithm will select no more than three masks per ROI.

Nuclei positions within each gland are extracted by the same DoG method as above. Using nuclear features of the gland, we wish to decide whether the gland comes from a cancerous ROI. Of the 453 ROI considered above, 353 have at least one gland detection; in all, 883 glands are detected. On average, one gland has on the order of a dozen nuclear detections.

The second part of Table 2 shows the results. Because the number of nuclei in the gland epithelium is much smaller than the number of nuclei in the ROI, the impact of a bad contour extraction is much greater. Accordingly, the positive impact of the HS method over the AC method is clearer than for ROI–based classification.

5 Discussion

We have introduced a novel contour extraction algorithm for radial boundaries that encloses regions by feature homogeneity, not edge strength. As a dynamic programming–based method with bounded radial jumps, it performs in linear time in the number of pixels to be searched for the boundary. In practice, this may be faster than using the active contour method with many initializations and parameter choices.

We have shown that its nuclear extractions in hematoxylin and eosin–stained gastric tissue are comparable or better to those of a well–tuned active contour edge–based contour extractor, in predicting cancer of an ROI. On gland classification, where a decision must be made with just a few nuclear extractions, it is clearly more powerful.

We believe that the homogeneity–based boundaries are better able to deal with unclear nuclear boundaries than edge–based boundaries. Dynamic programming allows a good global solution to be found without requiring an *a priori* decision of a smoothness versus local optimality tradeoff.

Training a classifier to accept or reject nuclear extractions, such as the one applied in the active contour method reviewed in this paper, could benefit the dynamic radial contour extractor as well. The comparisons in this paper compare the AC method with the benefit of such an advanced classifier, to the HS method without this benefit (simply filtering on luminance homogeneity).

6 Appendix: Proof of Theorem 1

For $i \in \{minRadius, \ldots, maxRadius\}$, let ρ_i denote the cheapest path to $(numTheta-1, i)$. In particular, we have $\rho_i(numTheta-1) = i$. Let $f(i) = \rho_i(0)$. If $|i - f(i)| \leq maxJump$ for some i, then the cheapest path to $(numTheta-1, i)$ may be extended to a path to $(numTheta, f(i))$ at finite cost, and the theorem holds.

Otherwise, $|i - f(i)| > maxJump$ for all i. In particular, $i \neq f(i)$ for all i. Then there exist $a < b$ with $f(a) > f(b)$; otherwise we would have

$$\begin{aligned} minRadius &< f(minRadius) < f(minRadius+1) < \cdots \\ &< f(maxRadius-1) < f(maxRadius) < maxRadius \end{aligned}$$

and there are not enough integers in the range from $minRadius$ to $maxRadius$ for all of these inequalities to hold simultaneously. Take such a and b.

Let s be the maximal value such that $\rho_a(\frac{2\pi s}{numTheta}) > \rho_b(\frac{2\pi s}{numTheta})$. Take $\alpha = \frac{2\pi s}{numTheta}$ and $\beta = \frac{2\pi(s+1)}{numTheta}$. Then $s < numTheta - 1$, $\rho_a(\alpha) > \rho_b(\alpha)$, and $\rho_a(\beta) \leq \rho_b(\beta)$.

Consequently,

$$\rho_a(\alpha) - \rho_b(\beta) \leq \rho_a(\alpha) - \rho_a(\beta) \leq maxJump \tag{8}$$

and

$$\rho_b(\beta) - \rho_a(\alpha) < \rho_b(\beta) - \rho_b(\alpha) \leq maxJump \tag{9}$$

so $|\rho_a(\alpha) - \rho_b(\beta)| \leq maxJump$ and $(s, \rho_a(\alpha))$ is a parent to $(s+1, \rho_b(\beta))$ in the Viterbi graph. Similarly,

$$\rho_b(\alpha) - \rho_a(\beta) < \rho_a(\alpha) - \rho_a(\beta) \leq maxJump \tag{10}$$

and

$$\rho_a(\beta) - \rho_b(\alpha) \leq \rho_b(\beta) - \rho_b(\alpha) \leq maxJump \tag{11}$$

so $(s, \rho_b(\alpha))$ is a parent to $(s+1, \rho_a(\beta))$ in the Viterbi graph. Let C_a be the total cost of the path up to $\rho_a(\alpha)$ and C_b be the total cost of the path up to $\rho_b(\alpha)$. If $C_a < C_b$, then a smaller–cost path to $(s+1, \rho_b(\beta))$ may be obtained by redefining ρ_b up to α to match ρ_a, contradicting the correctness of the Viterbi algorithm. If $C_a > C_b$, then redefining ρ_a up to α to match ρ_b provides a cheaper path, again a contradiction. Finally, if $C_a = C_b$, then ρ_a violated the convention of taking the lower–indexed parent, among parents of equal cost, at β. The theorem is proven. □

References

1. Montalto, M.C.: Pathology re-imagined: The history of digital radiology and the future of anatomic pathology. Arch. Path. and Lab. Medicine 132(5), 764–765 (2008)
2. Bloom, H.J., Richardson, W.W.: Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. Br. J. Cancer. 11, 359–377 (1957)
3. Nedzved, A., Ablameyko, S., Pitas, I.: Morphological segmentation of histology cell images. In: ICPR, vol. 1, pp. 500–503 (2000)
4. Latson, L., Sebek, B., Powell, K.A.: Automated cell nuclear segmentation in color images of hematoxylin and eosin-stained breast biopsy. Analytical and quantitative cytology and histology / the International Academy of Cytology (and) American Society of Cytology 25(6), 321–331 (2003)
5. Plissiti, M.E., Charchanti, A., Krikoni, O., Fotiadis, D.I.: Automated segmentation of cell nuclei in pap smear images. In: Proc. IEEE International Special Topic Conference on Information Technology in Biomedicine, Greece (2006)
6. Bamford, P., Lovell, B.C.: A methodology for quality control in cell nucleus segmentation. In: Digital Image Computing: Techniques and Applications (1999)
7. Pope, D.L., Parker, D.L., Clayton, P.D., Gustafson, D.E.: Left ventricular border recognition using a dynamic search algorithm. Radiology 155(2), 513–518 (1985)
8. Parker, D.L., Pryor, T.A.: Analysis of b-scan speckle reduction by resolution limited filtering. Ultrasonic Imaging 4(2), 108–125 (1982)
9. Sun, C., Pallottino, S.: Circular shortest path in images. Pattern Recognition 36(3), 709–719 (2003)

10. Appleton, B., Sun, C.: Circular shortest paths by branch and bound. Pattern Recognition 36(11), 2513–2520 (2003)
11. Adams, R., Bischof, L.: Seeded region growing. IEEE Trans. Pattern Anal. Mach. Intell. 16, 641–647 (1994)
12. Bellman, R.: Dynamic Programming. Princeton University Press, Princeton (1957)
13. Cosatto, E., Miller, M., Graf, H., Meyer, J.: Grading nuclear pleomorphism on histological micrographs. In: 19th International Conference on Pattern Recognition (2008)
14. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Prentice Hall, Englewood Cliffs (2002)
15. Japanese Research Society for Gastric Cancer. Japanese Classification of Gastric Carcinoma. Kanehara & Co., Ltd. (1995)
16. Malon, C., Marugame, A., Cosatto, E.: Epithelial structure detector (in preparation)

Robust Hyperplane Fitting Based on k-th Power Deviation and α-Quantile

Jun Fujiki[1], Shotaro Akaho[1], Hideitsu Hino[2], and Noboru Murata[2]

[1] National Institute of Advanced Industrial Science and Technology,
{jun-fujiki,s.akaho}@aist.go.jp
[2] Waseda University
hideitsu.hino@toki.waseda.jp, noboru.murata@eb.waseda.ac.jp

Abstract. In this paper, two methods for one-dimensional reduction of data by hyperplane fitting are proposed. One is least α-percentile of squares, which is an extension of least median of squares estimation and minimizes the α-percentile of squared Euclidean distance. The other is least k-th power deviation, which is an extension of least squares estimation and minimizes the k-th power deviation of squared Euclidean distance. Especially, for least k-th power deviation of $0 < k \leq 1$, it is proved that a useful property, called optimal sampling property, holds in one-dimensional reduction of data by hyperplane fitting. The optimal sampling property is that the global optimum for affine hyperplane fitting passes through N data points when an $N-1$-dimensional hyperplane is fitted to the N-dimensional data. The performance of the proposed methods is evaluated by line fitting to artificial data and a real image.

Keywords: hyperplane fitting, least k-th power deviations, least α-percentile of squares, optimal sampling property, random sampling.

1 Introduction

Dimensionality reduction is an important task in data processing, and *principal component analysis* (*PCA*) is, for example, commonly used to extract an essential structure of given data. PCA finds the subspace that maximizes the variance of the projected data. From the Pythagorean theorem, PCA is equivalent to the method of reducing the subspace that minimizes the variance of the projected data, which is called *minor component analysis* (*MCA*) [14]. In this paper, dimensionality reduction is considered in terms of reducing subspaces of data. Particularly, one-dimensional reduction is discussed throughout the paper.

In one-dimensional reduction of data, regression is one of the most important methods. In usual regression analysis, target variable is predicted by a linear combination of explanatory variables and the coefficients of the linear combination are estimated by minimizing residuals, which are the differences between true and predicted values of the target variable.

Least squares regression (*LS*), which is the most typical regression method, aims to minimize the sum of squared residuals. In *least absolute regression* [2] and

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 278–285, 2011.

Chebyshev regression [2], the sum of absolute residuals and the maximum of absolute residuals are minimized, respectively. In those regression models, it is assumed that all the explanatory variables do not contain any errors and only the target variable contains errors. There are other regression models in which the explanatory variables are assumed to be contaminated with errors. An example is *Deming regression* [3], or *orthogonal regression*, which minimizes sum of least mean squares of Euclidean distances between observed data and the fitted hyperplane. Deming regression is known to be equivalent to one-dimensional reduction of data by PCA, and is closely related to the *measurement error model* [1,10], which is recently investigated in statistics. Thus, hyperplane fitting by one-dimensional reduction, which is a kind of regression, is very important in data processing.

Also detecting lines in a two-dimensional image is an important example of one-dimensional reduction, and it is one of the most fundamental problems in image processing. To detect lines, the *Hough transformation* (*HT*) [4] and *random sampling consensus* (*RANSAC*) [7] are frequently used. The idea of HT is estimating the parameters of lines by voting on the cells in the parameter space. After the voting process, the parameters of lines are estimated as local maxima of the votes, and the data points are determined which line they belong to. In the HT, each data point is projected to a curve in the parameter space, i.e. many cells are voted by each data point, therefore a large computational cost is required to count the numbers of votes of all the cells in general. To overcome this drawback, the *randomized Hough transformation* (*RHT*) [13] is proposed. In the RHT, two data points are randomly sampled and a line that passes through those two points votes on the corresponding cell. Since each pair of points concerns with one cell only, the computational cost is drastically reduced. In the HT and RHT, the accuracy of detected lines depends on the resolution of cells in the parameter space. On the other hand, the idea of RANSAC is estimating the parameters of lines by counting the number of inliers. Many hyperplanes are iteratively generated by random sampling of the data points, and all the data are tested whether inlier or outlier with respect to the generated hyperplanes. In this test, each data point is regarded as an inlier when the distance between the data point and hyperplane is less than the given threshold. In RANSAC, the accuracy of detected lines depends on this threshold.

In this paper, two methods for one-dimensional reduction of data by hyperplane fitting are proposed. One is *least α-percentile of squares* (*L$_\alpha$PS*), which is an extension of *least median of squares* (*LMedS*) estimation [12]. The other is *least k-th power deviation* (*L$_k$PD*), which is an extension of LS estimation. Briefly speaking, L$_\alpha$PS uses the α-percentile instead of the median in LMedS, and L$_k$PD minimizes the sum of the k-th power deviations of errors instead of the sum of squares of errors. The regression models based on L$_k$PD for $1 \leq k \leq 2$ are already discussed in the previous papers [5,6,8,9,11]. In this paper, Deming regression by L$_k$PD, i.e. L$_k$PD of Euclidean distances for $0 < k \leq 1$, is considered. Especially, the *optimal sampling property* of L$_k$PD is discussed. This is a characteristic property that the global optimum for affine hyperplane fitting passes through N data points when an $N-1$ dimensional hyperplane is fitted to the N dimensional data, and it is proved that this property always holds in one-dimensional elimination of data by hyperplane fitting with L$_k$PD. Based on the optimal sampling property, an approximated optimization method for L$_k$PD is also proposed.

2 Hyperplane Fitting by Least k-th Power Deviations of Euclidean Distances

This section shows that (weighted) L$_k$PD of Euclidean distances has the optimal sampling property when $0 < k \leq 1$.

Let $\boldsymbol{x}$ be a variable in an N-dimensional space, and $\widetilde{\boldsymbol{x}} = (1, \boldsymbol{x}^\top)^\top$ be its homogeneous coordinate. D observations of the variable and their homogeneous coordinates are denoted by $\{\boldsymbol{x}_{[d]}\}_{d=1}^{D}$, and $\{\widetilde{\boldsymbol{x}}_{[d]}\}_{d=1}^{D}$, respectively. The equation of fitting affine hyperplane is written as

$$n_0 + \boldsymbol{n}^\top \boldsymbol{x} = \widetilde{\boldsymbol{n}}^\top \widetilde{\boldsymbol{x}} = 0 \quad \text{where} \quad ||\boldsymbol{n}|| = 1, \quad \widetilde{\boldsymbol{n}} = (n_0, \boldsymbol{n}^\top)^\top ,$$

and the Euclidean distance between the d-th observation $\boldsymbol{x}_{[d]}$ and the fitting hyperplane is denoted by $\widetilde{e}_{[d]} = |\widetilde{\boldsymbol{n}}^\top \widetilde{\boldsymbol{x}}_{[d]}|$ $(d = 1, \ldots, D)$. Then, weighted L$_k$PD is realized by minimizing the energy function

$$E_k = \sum_{d=1}^{D} w_{[d]} |\widetilde{\boldsymbol{n}}^\top \widetilde{\boldsymbol{x}}_{[d]}|^k = \sum_{d=1}^{D} \left| w_{[d]}^{\frac{1}{k}} \widetilde{\boldsymbol{n}}^\top \widetilde{\boldsymbol{x}}_{[d]} \right|^k$$
$$\text{subject to} \quad ||\boldsymbol{n}|| = 1, \quad w_{[d]} > 0 \, (d = 1, \ldots, D)$$

with respect to $\boldsymbol{n}$.

When the vector $\boldsymbol{n}$ belongs to an $N-1$-dimensional hypersphere, S^{N-1}, the vector $\widetilde{\boldsymbol{n}}$ belongs to an N-dimensional cylinder (Fig. 1), $\mathbb{R} \times S^{N-1}$, which is a convex hypersurface named $\widetilde{\mathcal{Q}}$ in this paper. The quadratic hypersurface $\widetilde{\mathcal{Q}}$ is divided to convex regions by D kinds of hyperplanes $\widetilde{\boldsymbol{x}}_{[d]}^\top \widetilde{\boldsymbol{n}} = 0$, and these

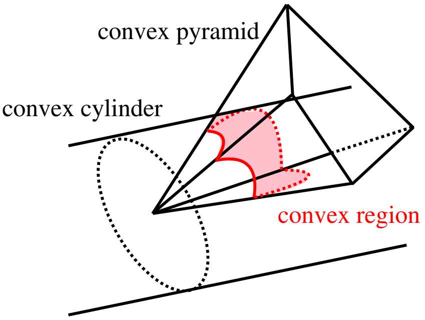


Fig. 1. The intersection between convex cylinder and a pyramid

hyperplanes make hyperpyramids. To distinguish each hyperpyramid, the sign of the hyperplane, which corresponds to the positive or negative region, is used. By defining the sign of each hyperplane as

$$s_{[d]} = \begin{cases} 1 & (\text{positive region of } \boldsymbol{x}_{[d]}^{\top}\boldsymbol{n} = 0) \\ -1 & (\text{negative region of } \boldsymbol{x}_{[d]}^{\top}\boldsymbol{n} = 0) \end{cases},$$

each hyperpyramid can be distinguished by the sign vector $\boldsymbol{s} = (s_1, \ldots, s_D)$, each hyperpyramid is denoted by $\mathcal{P}(\boldsymbol{s})$ in this paper. Also, each divided region of the quadratic hypersurface $\widetilde{\mathcal{Q}}$ is distinguished by $\boldsymbol{s}$ as the intersection of $\widetilde{\mathcal{Q}}$ and $\mathcal{P}(\boldsymbol{s})$, and each divided region is denoted by $\mathcal{D}(\boldsymbol{s})$. Because both of the quadratic hypersurface $\widetilde{\mathcal{Q}}$ and one of the hyperpyramid $\mathcal{P}(\boldsymbol{s})$ are convex regions, the region $\mathcal{D}(\boldsymbol{s})$, which is the intersection of $\widetilde{\mathcal{Q}}$ and $\mathcal{P}(\boldsymbol{s})$, is also a convex region.

Let a linear transformation $\mathcal{L}$ be

$$\mathcal{L} = \left(s_{[1]} w_{[1]}^{\frac{1}{k}} \widetilde{\boldsymbol{x}}_{[1]} \cdots s_{[D]} w_{[D]}^{\frac{1}{k}} \widetilde{\boldsymbol{x}}_{[D]} \right)^{\top} \in \mathbb{R}^{D \times (N+1)}$$

and the image of $\boldsymbol{n}$ by the linear transformation $\mathcal{L}$ be $\boldsymbol{\chi} = (\chi_{[1]}, \cdots, \chi_{[D]})^{\top} = \mathcal{L}\widetilde{\boldsymbol{n}} \in \mathbb{R}_{+}^{D}$. Then, there holds

$$E_k = \sum_{d=1}^{D} |\chi_{[d]}|^k = ||\boldsymbol{\chi}||_{L_k}^k ,$$

that is, E_k is the k-th power of L_k-norm of the vector $\boldsymbol{\chi}$. Since the vector $\boldsymbol{\chi}$ is on the image of the convex surface $\widetilde{\mathcal{Q}}$ by the linear transformation $\mathcal{L}$, $\boldsymbol{\chi}$ also belongs to the surface of some convex region. Therefore, $\mathcal{L}(\mathcal{D}(\boldsymbol{s}))$, which is the image of $\mathcal{D}(\boldsymbol{s})$ by the transformation $\mathcal{L}$, is the intersection of the convex quadratic hypersurface and pyramid $\mathcal{L}(\mathcal{P}(\boldsymbol{s}))$. Then, desired $\boldsymbol{n}$ is the minimizer of the L_k-distance between the origin of the coordinate and the point on the region $\mathcal{L}(\mathcal{D}(\boldsymbol{s}))$.

On $\mathbb{R}_{+}^{D}$, the set of the point within constant L_k-distance from the origin of the coordinate is concave when $0 < k < 1$, the minimum of the L_k-distance from the origin of the coordinate to the convex region $\mathcal{L}(\mathcal{D}(\boldsymbol{s}))$ is on the boundary of $\mathcal{L}(\mathcal{D}(\boldsymbol{s}))$, which is denoted by $\partial\mathcal{L}(\mathcal{D}(\boldsymbol{s}))$. Because the vector on the boundary $\partial\mathcal{L}(\mathcal{D}(\boldsymbol{s}))$ is on one of the hyperplanes of the pyramid, the minimum of the L_k-distance from the origin to the convex region $\partial\mathcal{L}(\mathcal{D}(\boldsymbol{s}))$ is on the boundary of $\partial\mathcal{L}(\mathcal{D}(\boldsymbol{s}))$, which is denoted by $\partial^2\mathcal{L}(\mathcal{D}(\boldsymbol{s}))$. By repeating this procedure, the minimum of the L_k-distance from the origin to the convex region $\mathcal{L}(\mathcal{D}(\boldsymbol{s}))$ is attained at one of the vertex of $\mathcal{L}(\mathcal{D}(\boldsymbol{s}))$. Because at least one vertex on $\mathcal{D}(\boldsymbol{s})$ is projected to the vertex of $\mathcal{L}(\mathcal{D}(\boldsymbol{s}))$ which gives the minimum, then the minimum of E_k is attained at least one of the vertex on $\mathcal{D}(\boldsymbol{s})$. This argument is true for all $\mathcal{D}(\boldsymbol{s})$'s, then the global optimum is attained at least one of the vertex of the set of $\mathcal{D}(\boldsymbol{s})$.

Here, the vertex of $\mathcal{D}(\boldsymbol{s})$ is the intersection of N hyperplanes, then the estimated hyperplane passes through N observed points, that is, affine hyperplane fitting by weighted L$_k$PD $(0 < k \leq 1)$ has the optimal sampling property.

3 Random Sampling Approximation

When the hyperplane fitting criterion satisfies the optimal sampling property, the fitting problem is reduced to combinatorial optimization of polynomial order, that is, we can find the optimum hyperplane among ${}_D\mathtt{C}_N$ $(< D^N)$ or ${}_D\mathtt{C}_{N-1}$ $(< D^{N-1})$ samplings in $N-1$-dimensional hyperplane fitting for D points. However, the number of combinations is quite large. Even for $D = 100$ and $N = 3$, it becomes about one million combinations, for example. To reduce the computational time, the random sampling technique can be adopted in order to approximate sampling all the combinations.

RANSAC and LMedS are typical methods utilizing random sampling. To find the optimum hyperplane, RANSAC tries to generate many hyperplanes by random sampling of the data points, and each data point is checked whether an inlier or an outlier of the generated hyperplanes. A data point is regarded as an inlier when the distance between the data point and the hyperplane is less than the given threshold e, and otherwise it is regarded as an outlier. Then, the optimum hyperplane is estimated as a hyperplane which has the maximum number of inliers. On the other hand, LMedS minimizes the median of the distances between data points and the hyperplane. LMedS does not require any predefined parameter such as a threshold in RANSAC, while we cannot use LMedS when the ratio of outliers is over 0.5 because the breakdown point of LMedS is 0.5 (50%). Therefore, in the case where the ratio of outliers is over 0.5, the median used in LMedS have to be extended to the α-percentile, which improves the breakdown point to $(100-\alpha)$%. This extension is called *$L_\alpha PS$*, (*least α-percentile of squares*) in this paper. In L$_\alpha$PS, the value α can be regarded as a parameter, which is an estimate of inlier-ratio. By the definition of L$_\alpha$PS, LMedS is represented as least 50-percentile squares (we denotes 50-L$_\alpha$PS in this paper), and the minimax criterion is represented as 100-L$_\alpha$PS. The major difference between RANSAC and L$_\alpha$PS is that RANSAC ignores the data which have larger errors than the given threshold, and L$_\alpha$PS ignores top $(100-\alpha)$% data with large errors.

Differently from these two methods, we propose a method which reduces the influence (weight) of data with large errors. In L$_k$PD, this is achieved by making the power k small, and particularly k is chosen as nonnegative and less than 2 $(0 \leq k < 2)$. Note that making the power k smaller in L$_k$PD corresponds to making the threshold smaller in RANSAC and making the percentile α smaller in L$_\alpha$PS. To illustrate the property of L$_\alpha$PS and L$_k$PD, numerical experiments are carried out in the next section.

4 Experiments

It is known that line fitting based on L_1-norm is more robust than usual line fitting based on L_2-norm. However, the estimated line based on L_1-norm sometimes derives bad result because of leverage point [12]. Then to reduce this effect by leverage point, 0.5-L$_k$PD is applied for line fitting, for example. Figure 2 shows

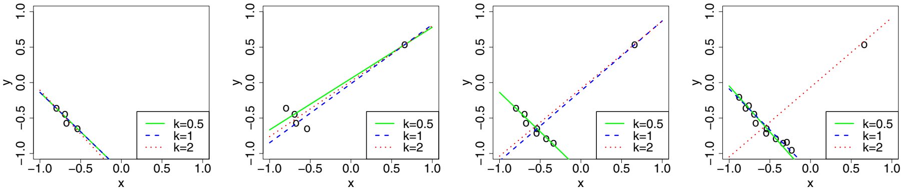


Fig. 2. Fitting results for 0.5-L$_k$PD, 1-L$_k$PD and 2-L$_k$PD (LS)

line fitting for four inliers, four inliers with one outlier, seven inliers with one outlier, and 11 inliers with one outlier, from left to right. In Fig. 2, 0.5-L$_k$PD, 1-L$_k$PD and 2-L$_k$PD (LS) are denoted as red dotted line, blue dashed line and green solid line, respectively. From Fig. 2, 0.5-L$_k$PD reduces the effect by leverage point more than 1-L$_k$PD. From this experiment, making k smaller is effective for robust estimation.

Figure 3 shows the relation of line fitting among RANSAC (cyan dotted line), 0.01-L$_k$PD(blue solid line), 0.5-L$_k$PD (blue dashed line), 5-L$_\alpha$PS(green solid line), LMedS (green dashed line) and LS (red dotted line). As shown in Fig. 3, we generated 70 data points in a stepwise manner and 20 points from the Gaussian distribution of its mean $(130, 70)^\top$ and variance $20.0\mathtt{I}_2$, where $\mathtt{I}_2$ is the two-dimensional identity matrix. As seen from Fig. 3, L$_k$PD and L$_\alpha$PS with appropriate k and α can reduce the effect of outliers and detect overall trend of the data. It is also seen that when k and α are set to very small values, these methods can detect fine structure from the contaminated observations.

Figure 4 shows the extracting line segments by RANSAC, 5-L$_\alpha$PS and 0.01-L$_k$PD. The original 640 × 480 image is converted to an edge image by Canny filter with Gaussian convolution function of $\sigma = 2.0$, and 4769 feature points having the peak value in the edge image are chosen. In this paper, line segment is assumed to consist of at least 20 points. When one line segment is estimated, the observed points of which distance from estimated line within $\sqrt{5}$-pixel is regarded

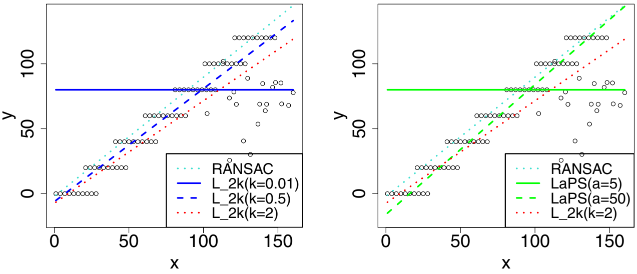


Fig. 3. Lines extraction for synthesized data: RANSAC (cyan), L$_k$PD (blue), L$_\alpha$PS (green) and LS (red)

as inlier and removed. Then the same procedure is iteratively applied to the rest of the observed points till no line segment is estimated. In the procedure, the number of random trials is determined as follows: When there are n points, the number of random trials such that at least one line is passing through two inliers among 20 inliers in probability $1 - 10^{-4}$ is approximated by $\dfrac{\log(10^{-4})}{\log\left\{1 - \dfrac{{}_{20}\mathrm{C}_2}{{}_{n}\mathrm{C}_2}\right\}}$.

Figure 4 shows the lines with more than 100 inliers among the extracted lines. As argued in section 3, RANSAC, LkPD, and LαPS are similar in that they are developed to reduce the effect of outliers. From an experimental result with a real-image, most of line segments in the picture are detected by either methods. As seen from Fig. 4 (middle), LαPS is suitable for line segments estimation and it can be used instead of RANSAC or HT. It is noted that, in this experiment, LkPD does not give the best performance. The reason why the result of LkPD is slightly inferior to others is that LkPD considers the effect of all points, while RANSAC and LαPS only consider the effect of the points around line segments.

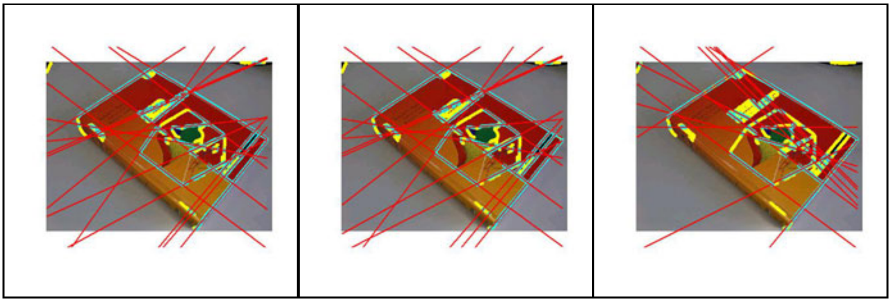


Fig. 4. Comparison of lines extraction methods: RANSAC (left), 5-LαPS (middle) and 0.01-LkPD (right)

5 Conclusion

In this paper, two methods for one-dimensional elimination of data by hyperplane fitting are proposed: one is least k-th power deviation (LkPD) and the other is least α-percentile of squares (LαPS). It is proved that in least k-th power deviation of $0 < k \leq 1$, the optimum $N-1$-dimensional hyperplane in an N-dimensional space is always represented by N data points, which is called the optimal sampling property. Based on this property, an optimization method for LkPD with random sampling is also presented.

Numerical experiments with simulated data show that LkPD and LαPS are considerably robust against outliers with appropriate values of the parameters, k and α, respectively. Also it is shown that LkPD and LαPS can successfully extract both local structure and global structure depending on the values of those parameters. In line extraction experiments with a real image, LαPS is comparable to RANSAC or HT, however the performance of LkPD is slightly inferior as compared to other two methods. This is because RANSAC and L$_\alpha$PS only consider the points around the target line segment, but L$_k$PD is affected by all the points. Hence, L$_k$PD is more suitable for extracting global or principal structure of data than local structure. For example, in practical application of epipolar geomety, estimation of the fundamental matrix is an important problem, and it is solved by the eight-point algorithm which is regarded as hyperplane fitting. This is an problem of extracting global structure, and a study on application of L$_k$PD to such a field remains as our future work.

References

1. Amari, S., Kawanabe, M.: Information geometry of estimating functions in semiparametric statistical models. Bernoulli 3(1), 29–54 (1997)
2. Appa, G., Smith, C.: On L_1 and Chebyshef estimation. Mathematical Programming 5(1), 73–78 (1973)
3. Deming, W.E.: Statistical adjustment of data. Wiley, NY (1943); Dover Publications edn. (1985)
4. Duda, R.O., Hart, P.E.: Use of the Hough transformation to detect lines and curves in pictures. Comm. ACM 15, 11–15 (1972)
5. Ekblom, H.: Calculation of linear best L_p-approximations. Bit Numerical Mathematics 13, 292–300 (1973)
6. Ekblom, H.: L_p-methods for robust regression. BIT Numerical Mathematics 14, 22–32 (1974)
7. Fischer, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Comm. ACM 24, 381–395 (1981)
8. Forsythe, A.B.: Robust estimation of straight line regression coefficients by minimizing p-th power deviations. Technometrics 14, 159–166 (1972)
9. Gentleman, W.M.: Robust estimation of multivariate location by minimizing p-th power deviations, Thesis at Princeton Univ., and Memorandum MM 65-1215-16, Bell Tel. Labs (1965)
10. Iba, Y., Akaho, S.: Gaussian process regression with measurement error. IEICE Trans. on Information and Systems E93-D(10), 2680–2689 (2010)
11. Rey, W.: On least p-th power methods in multiple regression and location estimations. Bit Numerical Mathematics 15(2), 174–184 (1975)
12. Rousseeuw, R.J., Leroy, A.M.: Robust Regression and Outlier Detection. John Wiley & Sons, NY (1987)
13. Xu, L., Oja, E., Kultanan, P.: A new curve detection method: Randomized Hough Transform (RHT). Pattern Recognition Letters 11(5), 331–338 (1990)
14. Xu, L., Oja, E., Suen, C.: Modified Hebbian learning for curve and surface fitting. Neural Networks 5(3), 441–457 (1992)

Incremental-Decremental Algorithm for Computing AT-Models and Persistent Homology*

Rocio Gonzalez-Diaz[1], Adrian Ion[2,3],
Maria Jose Jimenez[1], and Regina Poyatos[1]

[1] Applied Math Dep. (I), School of Computer Engineering,
University of Seville
{rogodi,majiro,rpoyatos}@us.es
[2] Pattern Recognition and Image Processing Group,
Vienna University of Technology
ion@prip.tuwien.ac.at
[3] Institute of Science and Technology, Austria

Abstract. In this paper, we establish a correspondence between the incremental algorithm for computing AT-models [8,9] and the one for computing persistent homology [6,14,15]. We also present a decremental algorithm for computing AT-models that allows to extend the persistence computation to a wider setting. Finally, we show how to combine incremental and decremental techniques for persistent homology computation.

Keywords: Persistent homology, AT-model for computing homology, cell complex.

1 Introduction

Homology is a topological invariant i.e. it is a property of an object which does not change under continuous (elastic) transformations of the object. Homology characterizes "holes" in any dimension (e.g. connected components, tunnels, cavities, etc.) by means of *cycles*. Given a combinatorial object made up by basic building blocks called *cells* (vertices, edges, faces, etc.), a cycle is a set of cells that "surround" a hole or a part of the object (e.g. a closed path in 2D, a closed path or a closed surface in 3D). Intuitively a *homology class* collects all cycles that "surround" the same hole (a precise mathematical definition is given later). The homology of a given object is then fully characterized by a basis of independent homology classes, which in turn is characterized by identifying one cycle, called *representative cycle*, for each of these classes.

Persistent homology studies homology classes and their lifetimes (persistence). Notice that while homology characterizes an object, persistent homology characterizes a sequence of growing object-instances i.e. an object together with an ordering for the cells (called a *filtration*). In recent years, persistent homology has found its way to applications, where it is mainly used to identify salient

* Partially supported by the Austrian Science Fund under P20134-N13.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 286–293, 2011.

features of an object in the presence of noise. E.g. find relevant local maxima without smoothing, compute the similarity of two objects as the similarity of their persistence information [5,3]. What all these applications have in common is that the object under study is fixed (e.g. one picture [3], one set of 3D sample points [4], one scan of a bone [7], etc.).

Current sensor and recording technologies provides not just one such recording but whole sequences over the temporal domain. Video has become ubiquitous, with decent to good quality recordings being produced by most mobile devices (phones, PDAs) and low priced webcams. Medical imaging is also moving from single 2D/3D image capture to recordings over a certain time period (e.g. sequences of ultrasound images). One way to deal with sequences is to take each frame separately and do all computation independently. However, due to temporal continuity, the same object can look very similar in consecutive frames and the overlap in the image can be high. Moreover, temporal information present in the sequence can help to identify salient features, as ones having a long lifetime not just over the filtration in the same frame, but also over multiple frames. There is no guarantee of inclusion or growing of objects in consecutive frames, some parts might "disappear" and some might "appear". A theory of decremental persistence, and incremental-decremental algorithms are needed for such arbitrary changes in the object.

In this paper we describe a further step in providing such a theory. First, we establish a correspondence between the incremental algorithm for computing AT-models [8,9] and the one for computing persistent homology [6,14,15]. Then, we provide a decremental algorithm for computing AT-models, suitable for extending the computation of persistence with the combination of an incremental-decremental technique.

2 Background

We consider $\mathbf{Z}/2$ as the ground ring throughout the paper.

Roughly speaking, a *cell complex* is a general topological structure by which a space is decomposed into basic elements (*cells*) of different dimensions, which are *glued* together by their boundaries (see a formal definition of CW-complex in [12]). If the building blocks (cells) of a cell complex are convex polytopes (vertices, edges, polygons, polyhedra, ...) then the cell complex is a *polyhedral cell complex.* Given a (polyhedral) cell complex K, a *proper face* of $\sigma \in K$ is a face of σ whose dimension is strictly less than the one of σ. A *facet* of σ is a proper face of σ of maximal dimension. A *maximal cell* of K is a cell of K which is not a proper face of any other cell of K.

For any graded set $S = \{S_p\}_p$ (subscript is used to denote the dimension of the elements), one can consider formal sums of elements of $S_p = \{s_p^1, \dots, s_p^{m_p}\}$, for a fixed p, which are called *p-chains*, and which form an abelian group, denoted by $C_p(S)$, with respect to the component-wise addition (mod 2). Therefore, a p-chain c is $c = \sum_{i=1}^{m} a^i s_p^i$, where $a^i \in \mathbf{Z}/2$ for $i = 1, ..., m$. This way, $s_p^i \in c$ if $a^i = 1$. The collection of all the chain groups associated to S, $\{C_p(S)\}_p$, is called also chain group, for simplicity. A *chain complex* is a collection $\mathcal{C}(S) = \{C_p(S), \partial_p^S\}_p$, where $\partial^S = \{\partial_p^S : C_p(S) \to C_{p-1}(S)\}$ is a square zero homomorphism (i.e., $\partial_{p-1}\partial_p \equiv 0$) called the *boundary operator*. The boundary of a q-cell is the formal sum of all its facets. It is extended to q-chains by linearity. A homomorphism $f = \{f_p : C_p(S) \to C_p(S')\}_p$ is a *chain map* if $f_{p-1}\partial_p^S \equiv \partial_p^{S'} f_p$, for all p. For simplicity, we sometimes write $f : C(S) \to C(S')$ instead of $f = \{f_p : C_p(S) \to C_p(S')\}_p$ and $f(\sigma)$ instead of $f_p(\sigma)$. A p-chain $a \in C_p(S)$ is called a *p-cycle* if $\partial_p^S a = 0$. If $a = \partial_{p+1}^S b$ for some $b \in C_{p+1}(S)$ then a is called a *p-boundary.* We say that two p-cycles a and b are *homologous* if there exists a $(p+1)$-chain $c \in C_{p+1}(S)$ such that $a = b + \partial_{p+1}^S c$. Define the *p-th homology group* to be the quotient group of p-cycles mod p-boundaries denoted by $H_p(S)$. Each element $[a]$ of $H_p(S)$ is a quotient class obtained by adding each p-boundary to a given p-cycle a called a *representative cycle* of the homology class $[a]$. The *homology* of S is the chain group $\mathcal{H}(S) = \{H_p(S)\}_p$. See [13] for further details.

3 AT-Model

An algebraic topological (AT) model (implicitly used in [8] and first defined in [9]) for a given cell complex K not only permits to compute homology but also finer topological invariants such as cohomology or the cohomology ring.

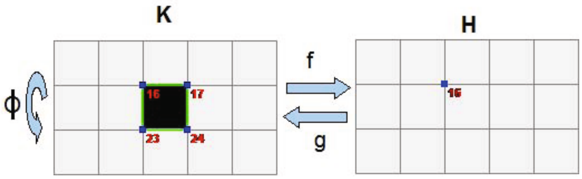
An *AT-model* for a cell complex K is an algebraic set (f, g, ϕ, K, H), where:

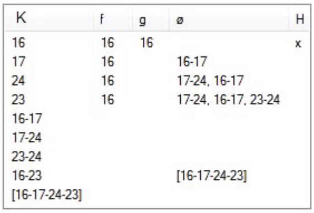
- K is the cell complex.
- $H \subseteq K$ describes the homology of K, in the sense that it contains a distinct p-cell for each p-homology class of a basis, for all p. The cells in H are called *surviving cells.* Since $\partial_p^H \equiv 0$ for all p, $\mathcal{C}(H)$ is simply the chain group $\{C_p(H)\}_p$. Moreover, $\mathcal{C}(H)$ (the chain group generated by H) and $\mathcal{H}(K)$ (the homology of K) are isomorphic [8,9]. Therefore, every cell of H corresponds to a homology class generator.
- $g = \{g_p : C_p(H) \to C_p(K)\}_p$ is a chain map that maps each p-cell h in H to one representative cycle $g_p(h)$ of the corresponding class $[g_p(h)]$ in $H_p(K)$.
- $f = \{f_p : C_p(K) \to C_p(K)\}_p$ is a chain map that maps each p-cell x in K to a sum of surviving cells, satisfying that if $a, b \in C_p(K)$ are two homologous p-cycles then $f_p(a) = f_p(b)$. Moreover, $fg(x) = x$ for any $x \in H$.
- $\phi = \{\phi_p : C_p(K) \to C_{p+1}(K)\}_p$ is a *chain homotopy* (see [13]) from gf to the identity homomorphism of $\mathcal{C}(K)$. Intuitively $\phi(\sigma)$ returns the cells needed to be contracted to "bring" σ to a surviving cell.

Fig. 1 is an example of an AT-model for a single pixel codified as a cubical complex.

In order to establish a connection between the existing algorithms for computing AT-models [9] and the theory of persistent homology [6,14,15], we add two **extra-conditions** which ensure that the basis for $\mathcal{H}(K^i)$ is maintained implicitly through the cells in H^i (see Section 5 for more details):

(P1) Annihilation: $\phi_{p+1}\phi_p \equiv 0$, $f_{p+1}\phi_p \equiv 0$ and $\phi_p g_p \equiv 0$.
(P2) If $h \in K$ is a surviving p-cell then $f_p(h) = h$ and $\phi_p(h) = 0$.





K	f	g	ø	H
16	16	16		x
17	16		16-17	
24	16		17-24, 16-17	
23	16		17-24, 16-17, 23-24	
16-17				
17-24				
23-24				
16-23			[16-17-24-23]	
[16-17-24-23]				

Fig. 1. AT-model for a single pixel

Lemma 1. *The extra-conditions (P1) and (P2) are satisfied by the incremental and decremental algorithm for computing AT-models presented in this paper.*

4 Algorithms for Computing AT-Models

Now, we give an intuitive approach to the incremental algorithm for computing AT-models given in [8] in order to establish an interpretation in terms of persistence, and present a decremental algorithm for computing AT-models with the aim of extending the computation of persistent homology to a more general setting.

Incremental AT-model: Let K be a cell complex with a full ordering of its cells, $\{\sigma^1, \dots, \sigma^n\}$, satisfying that if σ^i is a face of σ^j then $i < j$. Consider a filtration of K, i.e. a nested sequence of subcomplexes, $\emptyset = K^0 \subseteq K^1 \subseteq K^2 \subseteq \cdots \subseteq K^n$ such that $K^i = \{\sigma^1, \dots, \sigma^i\}$. All the proper faces of σ^i are in K^{i-1}.

First, define an AT-model $(f^1, g^1, \phi^1, K^1, H^1)$ for K^1: $H^1 := \{\sigma^1\}$, $f^1(\sigma^1) := \sigma^1$, $g^1(\sigma^1) := \sigma^1$ and $\phi^1(\sigma^1) := 0$. Second, successively add a new cell σ^i, for $i = 2, \dots, n$, computing a new AT-model $(f^i, g^i, \phi^i, K^i, H^i)$ for $K^i = K^{i-1} \cup \{\sigma^i\}$, as follows: Initially, $H^i := H^{i-1}$; $f^i(\mu) := f^{i-1}(\mu)$ and $\phi^i(\mu) := \phi^{i-1}(\mu)$, for any $\mu \in K^i$ different from σ^i; $g^i(h) := g^{i-1}(h)$, for any $h \in H^i$. Consider $f^{i-1}\partial(\sigma^i)$ to detect if σ^i will create or destroy a homology class:

(1) If $f^{i-1}\partial(\sigma^i) = 0$ (a new homology class is created) then, $H^i := H^{i-1} \cup \{\sigma^i\}$, $f^i(\sigma^i) := \sigma^i$, $g^i(\sigma^i) := \sigma^i + \phi^{i-1}\partial(\sigma^i)$ and $\phi^i(\sigma^i) := 0$.

(2) If $f^{i-1}\partial(\sigma^i) \neq 0$ (a homology class is destroyed), let j be the largest index such that $\sigma^j \in f^{i-1}\partial(\sigma^i)$. Then, observe that $j < i$ and dim σ^j =dim $\sigma^i - 1$. Then, $H^i := H^{i-1} \setminus \{\sigma^j\}$, $f^i(\sigma^i) := 0$, $\phi^i(\sigma^i) := 0$.
Besides, two operations are applied to all cells $x \in K^i$ such that $\sigma^j \in f^{i-1}(x)$:
   - Update f: $f^i(x) := f^{i-1}(x) + f^{i-1}\partial(\sigma^i)$. Intuitively, it "propagates" over σ^i the information for f of the cells in $\partial\sigma^i$ and cancels out σ^j.
   - Update ϕ: $\phi^i(x) := \phi^{i-1}(x) + \sigma^i + \phi^{i-1}\partial(\sigma^i)$. Roughly speaking, $\sigma^i + \phi^{i-1}\partial(\sigma^i)$ is a connection between the old and the new surviving cell(s).

Relation to persistent homology: The algorithm for computing persistent homology that appears in [6,14,15], marks a k-cell σ^i as positive if it belongs to a k-cycle in $C(K^i)$ (σ^i creates a new homology class at time i) and negative otherwise (σ^i destroys the homology class created before by σ^j for $j < i$ and, in this case, σ^i is paired with σ^j). The following lemmas show the equivalence between these concepts and the incremental AT-model above.

Lemma 2. *σ^i belongs to a k-cycle in $C(K^i)$ if and only if $f^{i-1}\partial(\sigma^i) = 0$.*

Proof. If $f^{i-1}\partial(\sigma^i) = 0$, then $\partial\sigma^i = \partial\phi^{i-1}(\partial\sigma^i)$. Therefore, $\sigma^i + \phi^{i-1}\partial(\sigma^i)$ is a k-cycle and σ^i belongs to it. Conversely, if σ^i belongs to a k-cycle a in $C(K^i)$, then $a = \sigma^i + b$ where b is a k-chain in $C(K^{i-1})$. Since $\partial a = 0$ then $f^{i-1}(\partial\sigma^i) = f^{i-1}(\partial b)$. Since $f^{i-1}(\partial b) = \partial f^{i-1}(b) = 0$ then $f^{i-1}(\partial\sigma^i) = 0$. □

By Lemma 2, the fact of marking a cell σ^i as positive is equivalent to holding condition $f^{i-1}\partial(\sigma^i) = 0$ in the incremental algorithm for computing AT-models.

Following the theory of persistent homology, a *canonical cycle* c^i is a non-bounding cycle that contains σ^i but no other positive cell.

Lemma 3. *If σ^i is positive, then $g^i(\sigma^i) = \sigma^i + \phi^{i-1}\partial(\sigma^i)$ is a canonical cycle.*

At time i, the youngest cell $\sigma^j \in \Gamma(\partial\sigma^i)$ is paired with σ^i, identifying σ^i as the destroyer of the homology class created by σ^j. Once one has the pairing of positive and negative cells, computing the persistent Betti numbers is trivial. To measure the life-time of a non-bounding cycle, one has to find when the cycle's homology class is created by a positive cell and destroyed by a negative cell. To detect these events, the collection of positive k-cells $\Gamma(d)$ for a given cycle d such that d and $\sum_{\sigma^g \in \Gamma(d)} c^g$ are homologous, is obtained using the incremental method for computing AT-models as follows:

Lemma 4. *Any k-cycle d in $C(K^{i-1})$ is homologous to $g^{i-1}f^{i-1}(d)$. Moreover, $\Gamma(d) = f^{i-1}(d)$ and $g^{i-1}f^{i-1}(d) = \sum_{\sigma^g \in \Gamma(d)} c^g$.*

Decremental AT-model: Let (f, g, ϕ, K, H) be an AT-model for a cell complex K satisfying the extra-conditions (P1) and (P2). Let σ be a maximal cell of K. Then an AT-model for $K' = K \setminus \{\sigma\}$ is constructed as follows:

Algorithm 1. *Initially, $H' := H$, $g'(h) := g(h)$ for all $h \in H'$, $f'(x) := f(x)$ and $\phi'(x) := \phi(x)$ for all $x \in K'$.*

(1) If there exists $\beta \in H$ such that $\sigma \in g(\beta)$ then σ destroys the homology class $[g(\beta)]$ created before by β. Therefore, $H' := H \setminus \{\beta\}$; $f'(x) := f(x) + \beta$ if $\beta \in f(x)$ and $x \in K'$; $g'(h) := g(h) + g(\beta)$ if $\sigma \in g(h)$ and $h \in H'$; $\phi'(y) := \phi(y) + g(\beta)$ if $\sigma \in \phi(y)$ and $y \in K'$.

(2) Otherwise, since $\sigma \in \phi\partial\sigma$, there exists $\mu \in \partial\sigma$, $\mu \notin H$, such that $\sigma \in \phi(\mu)$. Then μ creates a new homology class $[g'(\mu)]$. Therefore: $H' := H \cup \{\mu\}$; $g'(\mu) := gf(\mu) + \partial\phi(\mu)$. $f'(x) := f(x) + \mu + f(\mu)$ and $\phi'(x) := \phi(x) + \phi(\mu)$ if $\sigma \in \phi(x)$ and $x \in K'$.

In order to satisfy the following proposition, the formulas for f' and ϕ' in step (2) above are different to the ones given in [9].

Proposition 1. *The output of Alg. 1, (f', g', ϕ', K', H'), is an AT-model for $K' = K \setminus \{\sigma\}$ satisfying the extra-conditions (P1) and (P2).*

Proof. Proof of step (1) in Alg. 1 is given in [9]. The verification of the rest of the properties follows a similar strategy and is left to the reader. □

5 Incremental-Decremental Algorithm for Computing AT-Models and Persistent Homology

Now, let $\emptyset = K^0 \leftrightarrow K^1 \leftrightarrow \cdots \leftrightarrow K^n$ be a sequence of cell complexes (that we call *a zig-zag filtration*), such that every two consecutive complexes differ by a single cell, i.e. either $K^i = K^{i-1} \cup \{\sigma\}$ or $K^i = K^{i-1} \setminus \{\sigma\}$. Let $\{\sigma^1, \dots, \sigma^m\}$, $m \leq n$, be the ordered set of all the cells added in a given zig-zag filtration such that if $i < j$ then σ^i was added before σ^j to the filtration. Then, one can consider the sequence of homology groups $H(K^0) \leftrightarrow H(K^1) \leftrightarrow \cdots \leftrightarrow H(K^n)$ where the connecting homomorphisms are induced by inclusion.

Incremental-decremental algorithm: Initially, $H^1 := \{\sigma^1\}$, $f^1(\sigma^1) := \sigma^1$, $g^1(\sigma^1) := \sigma^1$ and $\phi^1(\sigma^1) := 0$. At time i, a cell σ is added or removed. Then, use the incremental or decremental algorithm presented here, respectively, for computing the AT-model $(f^i, g^i, \phi^i, K^i, H^i)$. Two cases can occur:

(1) A homology class is created by a positive cell μ. If $K^i = K^{i-1} \cup \{\sigma\}$ then $\mu := \sigma$. If $K^i = K^{i-1} \setminus \{\sigma\}$, then $\mu := \sigma^j$ where σ^j is the youngest cell in $\partial\sigma$ such that $\sigma \in \phi^{i-1}(\sigma^j)$. The cell μ is added to H^{i-1} to get H^i.

(2) A homology class represented by a positive cell σ^j is destroyed by a negative cell σ^k, $j < k \leq i$. If $K^i = K^{i-1} \cup \{\sigma\}$ then $\sigma^k := \sigma$ and σ^j is the youngest cell in $f^{i-1}\partial(\sigma)$. If $K^i = K^{i-1} \setminus \{\sigma\}$, then $\sigma^k := \sigma$ and σ^j is the youngest cell in H^{i-1} such that $\sigma^k \in g^{i-1}(\sigma^j)$. The cell σ^j is removed from H^{i-1} to get H^i.

If a cell μ creates a homology class at time j and it is destroyed at time i, $j < i$, then a horizontal line from (j, ℓ) to (i, ℓ) is added to the corresponding barcode (see [2]); If a cell μ creates a homology class at time i and it survives along the process then a horizontal line from (i, ℓ) to (∞, ℓ) is added, where ℓ is the index of the cell μ in the given ordering of the cells. See examples of barcodes using incremental-decremental algorithm for computing AT-models in Fig. 2, 3, 4. See Fig. 5 as an example of application with digital images.

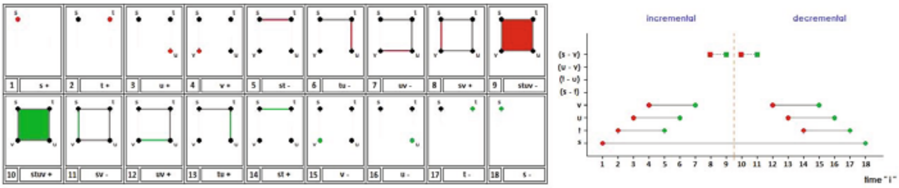


Fig. 2. Symmetric zig-zag filtration and barcode

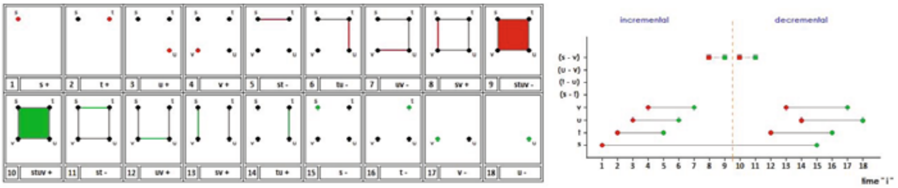


Fig. 3. Non-symmetric zig-zag filtration and barcode

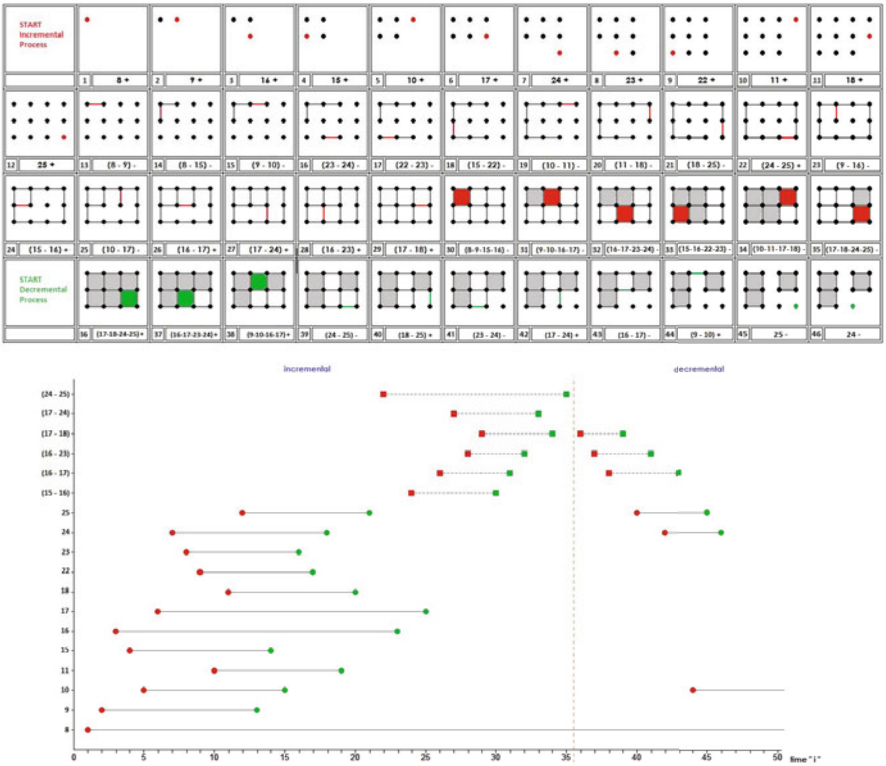


Fig. 4. Example of zig-zag filtration and barcode

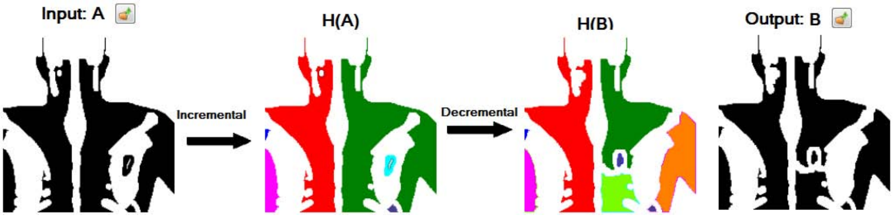


Fig. 5. Example of application. Each color refers to one connected component

6 Future Work

The proposed method is able to deal with a general filtration, allowing randomly adding or removing a cell. This is different from both standard persistence and zigzag persistence [2], which compute the filtration from a single scalar function. A correspondence between the algorithms presented here and the one given in [2] is left as a future work.

The presented algorithm is valid for any dimension but with $\mathbf{Z}/2$ domain; how can persistence for integer homology be defined? The results given in [10] may be used to try to find the answer. Since the computation of AT-models allows the computation of finer invariants than homology such as the cohomology ring [8], how could we deal with persistence of other (finer) topological invariants? We also plan to deal with the problem of extending persistence to other geometrical operations such as face removal, simplicial collapse and edge contractions using AT-models by means of the initial results given in [11].

References

1. Biasotti, S., Giorgi, D., Spagnuolo, M., Falcidieno, B.: Size functions for comparing 3D models. Pattern Recogn. 41(9), 2855–2873 (2008)
2. Carlsson, G., de Silva, V., Morozov, D.: Zigzag persistent homology and real-valued functions. In: SoCG 2009, pp. 247–256. ACM, New York (2009)
3. Chazal, F., Cohen-Steiner, D., Guibas, L., Mémoli, F., Oudot, S.: Gromov-Hausdorff Stable Signatures for Shapes using Persistence. Comput. Graph. Forum 28(5), 1393–1403 (2009)
4. Chazal, F., Guibas, L., Oudot, S., Skraba, P.: Analysis of scalar fields over point cloud data. In: Proc. of SODA 2009, pp. 1021–2030 (2009)
5. Cerri, A., Ferri, M., Giorgi, D.: Retrieval of trademark images by means of size functions. Graph. Models 68(5), 451–471 (2006)
6. Edelsbrunner, H., Letscher, D., Zomorodian, A.: Topological persistence and simplification. In: FOCS 2000, pp. 454–463. IEEE Computer Society, Los Alamitos (2000)
7. Ferri, M., Stanganelli, I.: Size functions for the morphological analysis of melanocytic lesions. Journal of Biomedical Imaging 5, 1–5 (2010)
8. Gonzalez-Diaz, R., Real, P.: On the cohomology of 3D digital images. Discrete Applied Math. 147(2-3), 245–263 (2005)
9. Gonzalez-Diaz, R., Medrano, B., Real, P., Sanchez-Pelaez, A.: Simplicial perturbation techniques and effective homology. In: Ganzha, V.G., Mayr, E.W., Vorozhtsov, E.V. (eds.) CASC 2006. LNCS, vol. 4194, pp. 166–177. Springer, Heidelberg (2006)
10. Gonzalez-Diaz, R., Jimenez, M., Medrano, B., Real, P.: A tool for integer homology computation: Lambda-AT model. Image and Vision Computing, 1–9 (2008)
11. Gonzalez-Diaz, R., Jimenez, M., Medrano, B., Molina, H., Real, P.: Integral operators for computing homology generators at any dimension. In: Ruiz-Shulcloper, J., Kropatsch, W.G. (eds.) CIARP 2008. LNCS, vol. 5197, pp. 356–363. Springer, Heidelberg (2008)
12. Hatcher, A.: Algebraic Topology. Cambridge University Press, Cambridge (2002)
13. Munkres, J.: Elements of Algebraic Topology. Addison-Wesley Co., Reading (1984)
14. Zomorodian, A.: Topology for Computing. Cambridge Monographs on Applied and Computational Mathematics (16) (2005)
15. Zomorodian, A., Carlsson, G.: Computing persistent homology. Discrete and Computational Geometry 33(2), 249–274 (2005)

Persistent Betti Numbers for a Noise Tolerant Shape-Based Approach to Image Retrieval

Patrizio Frosini[1,3] and Claudia Landi[2,3]

[1] Dipartimento di Matematica, Università di Bologna
frosini@dm.unibo.it
[2] DiSMI, Università di Modena e Reggio Emilia
clandi@unimore.it
[3] ARCES, Università di Bologna

Abstract. In content-based image retrieval a major problem is the presence of noisy shapes. It is well known that persistent Betti numbers are a shape descriptor that admits a dissimilarity distance, the matching distance, stable under continuous shape deformations. In this paper we focus on the problem of dealing with noise that changes the topology of the studied objects. We present a general method to turn persistent Betti numbers into stable descriptors also in the presence of topological changes. Retrieval tests on the Kimia-99 database show the effectiveness of the method.

Keywords: Multidimensional persistent homology, Hausdorff distance, symmetric difference distance.

1 Introduction

Persistence is a theory for studying objects related to computer vision and computer graphics, by adopting different functions (e.g., distance from the center of mass, distance from the medial axis, height, geodesic distance, color mapping) to measure the shape properties of the object under study (e.g., roundness, elongation, bumpiness, color). The object, considered as a topological space, is explored through the sequence of nested sub-level sets of the considered measuring function. A shape descriptor, called a persistent homology group, can be constructed by encoding at which scale a topological feature (e.g., a connected component, a tunnel, a void) is created, and when it is annihilated along this filtration. For application purposes, these groups are further encoded by considering only their dimension, yielding a parametrized version of Betti numbers, known in the literature as *persistent Betti numbers* [14], a *rank invariant* [8], and, for the 0th homology, a *size function* [19].

In the literature, a large number of methods for shape matching has been proposed, such has the shape-context [1], the shock graph [17], and the inner distance [5], to name a few. Persistent Betti numbers are shape descriptors belonging to the class of shape-from-functions methods which are widely reviewed in [4].

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 294–301, 2011.

The stability of persistent Betti numbers functions (hereafter, PBNs, for brevity) is quite an important issue, every data measurement being affected by noise. The stability problem involves both stability under perturbations of the topological space that represents the object, and stability under perturbations of the function that measures the shape properties of the object.

The problem of stability with respect to perturbations of the measuring function has been studied in [10] for scalar-valued measuring functions. For vector-valued measuring functions, the multidimensional matching distance between PBNs is introduced in [6], and is shown to provide stability in [9]. For the case of 0th homology, this problem is treated in [12] and [3] for scalar- and vector-valued functions, respectively.

In this paper we consider the problem of stability of PBNs with respect to changes of the topological space. This topic has been studied in [11] for sublevel sets of smooth functions satisfying certain conditions on the norm of the gradient. Unfortunately these conditions seem not to be satisfied in a wide variety of situations common in object recognition, such as point cloud data, curves in the plane, domains affected by salt & pepper noise.

We propose a general approach to the problem of stability of PBNs with respect to domain perturbations that applies to more general domains, i.e. compact subsets of $\mathbb{R}^n$. Moreover, according to the type of noise affecting the data, we propose to choose an appropriate set distance to measure the domain perturbation (for example, the Hausdorff distance in case of small position errors, the symmetric difference pseudo-distance in the presence of outliers). The core of our approach is to choose an appropriate continuous function to represent the domain, so that the problem of stability for noisy domains with respect to a given set distance can be reduced to that of stability with respect to changes of the functions. This is achieved by substituting the domain K with an appropriate function f_K defined on a fixed set X containing K. Assuming we were interested in the shape of K, as seen through a measuring function $\boldsymbol{\varphi}_{|K} : X \to \mathbb{R}^k$, we actually study the function $\boldsymbol{\Phi} : X \to \mathbb{R}^{k+1}$, with $\boldsymbol{\Phi} = (f_K, \boldsymbol{\varphi})$. Persistent Betti numbers of $\boldsymbol{\Phi}$ can be compared using the multidimensional matching distance, thus obtaining robustness of PBNs under domain perturbations.

In particular, we use this strategy when sets are compared by the Hausdorff distance and by the symmetric difference pseudo-distance. In both these cases we show stability results (Theorems 1 and 3). Moreover we show the relation existing between the shape of K as described by $\boldsymbol{\varphi}_{|K}$ and the shape described by $\boldsymbol{\Phi} = (f_K, \boldsymbol{\varphi})$ (Theorem 2).

Finally, we conclude our paper presenting some experiments in which our method is tested on the Kimia-99 database [16], using as query shapes noisy versions of the original shapes.

2 Preliminaries on PBNs

Persistence may be used to construct shape descriptors that capture both geometrical and topological properties of objects $K \subset \mathbb{R}^n$. Geometrical properties of K are studied through the choice of a function $\boldsymbol{\varphi} = (\varphi_i) : K \to \mathbb{R}^k$, each component φ_i describing a shape property. The function $\boldsymbol{\varphi}$ is called a k-dimensional *measuring* (or *filtering*) *function*. Topological properties of K as seen through $\boldsymbol{\varphi}$ are studied by considering sub-level sets $K\langle \boldsymbol{\varphi} \preceq \boldsymbol{u} \rangle = \{x \in K : \varphi_i(x) \le u_i,\ i = 1, \ldots, k\}$. For $\boldsymbol{u} = (u_i), \boldsymbol{v} = (v_i) \in \mathbb{R}^k$ with $u_i \le v_i$, (briefly, $\boldsymbol{u} \preceq \boldsymbol{v}$), the sub-level set $K\langle \boldsymbol{\varphi} \preceq \boldsymbol{u} \rangle$ is contained in the sub-level set $K\langle \boldsymbol{\varphi} \preceq \boldsymbol{v} \rangle$. A classical transform of algebraic topology, called homology, provides topological invariants. Working with homology coefficients in a field, it transforms topological spaces into vector spaces and continuous maps (e.g., inclusions) into linear maps. This leads to the following definition.

Definition 1 (Persistent Betti Numbers). *Let $q \in \mathbb{Z}$. Let $\pi_q^{(\boldsymbol{u},\boldsymbol{v})} : \check{H}_q(K\langle \boldsymbol{\varphi} \preceq \boldsymbol{u}\rangle) \to \check{H}_q(K\langle \boldsymbol{\varphi} \preceq \boldsymbol{v}\rangle)$ be the homomorphism induced by the inclusion map $\pi^{(\boldsymbol{u},\boldsymbol{v})} : K\langle \boldsymbol{\varphi} \preceq \boldsymbol{u}\rangle \hookrightarrow K\langle \boldsymbol{\varphi} \preceq \boldsymbol{v}\rangle$ with $\boldsymbol{u} \preceq \boldsymbol{v}$, where $\check{H}_q$ denotes the qth Čech homology group. The q*th persistent Betti number function *of $\boldsymbol{\varphi}$ is the function $\beta_{\boldsymbol{\varphi}} : \{(\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^k \times \mathbb{R}^k : \boldsymbol{u} \prec \boldsymbol{v}\} \to \mathbb{N} \cup \{\infty\}$ defined as $\beta_{\boldsymbol{\varphi}}(\boldsymbol{u}, \boldsymbol{v}) = \dim \operatorname{im} \pi_q^{(\boldsymbol{u},\boldsymbol{v})}$.*

If K is a triangulable space embedded in some $\mathbb{R}^n$, then $\beta_{\boldsymbol{\varphi}}(\boldsymbol{u}, \boldsymbol{v}) < +\infty$, for every $\boldsymbol{u} \prec \boldsymbol{v}$ and every $q \in \mathbb{Z}$ [7]. Clearly, $\boldsymbol{u} \prec \boldsymbol{v}$ means $u_i < v_i$ for $i = 1, \ldots, k$.

In order to get a dissimilarity measure between the shapes described by two PBNs, in the case of scalar-valued measuring functions, we can use the matching distance d_{match}, also known as the bottleneck distance between persistence diagrams [10]. In the case of vector-valued measuring functions, we can utilize the foliation method to obtain the following distance via a reduction to the case of scalar-valued measuring functions, as described in [6].

Definition 2. *The distance D_{match} between the PBNs of two vector-valued measuring functions $\boldsymbol{\varphi}, \boldsymbol{\psi} : K \to \mathbb{R}^k$ is defined as follows:*

$$D_{match}(\beta_{\boldsymbol{\varphi}}, \beta_{\boldsymbol{\psi}}) = \sup_{(\boldsymbol{l},\boldsymbol{b}) \in Adm_k} \min_i l_i \cdot d_{match}\left(\beta_{F^{\boldsymbol{\varphi}}_{(\boldsymbol{l},\boldsymbol{b})}}, \beta_{G^{\boldsymbol{\psi}}_{(\boldsymbol{l},\boldsymbol{b})}}\right),$$

where $Adm_k = \{(\boldsymbol{l}, \boldsymbol{b}) \in \mathbb{R}^k \times \mathbb{R}^k : l_i > 0, \sum_{i=1}^k l_i^2 = 1, \sum_{i=1}^k b_i = 0\}$, $F^{\boldsymbol{\varphi}}_{(\boldsymbol{l},\boldsymbol{b})} : K \to \mathbb{R}$, $F^{\boldsymbol{\varphi}}_{(\boldsymbol{l},\boldsymbol{b})}(x) = \max_{i=1,\ldots,k}\left\{\frac{\varphi_i(x)-b_i}{l_i}\right\}$, and $G^{\boldsymbol{\psi}}_{(\boldsymbol{l},\boldsymbol{b})} : K \to \mathbb{R}$, $G^{\boldsymbol{\psi}}_{(\boldsymbol{l},\boldsymbol{b})}(x) = \max_{i=1,\ldots,k}\left\{\frac{\psi_i(x)-b_i}{l_i}\right\}$.

The key property of D_{match} is that it inherits stability with respect to perturbations of the measuring function from the stability of d_{match} assuming $\boldsymbol{\varphi}, \boldsymbol{\psi}$ only continuous [9].

3 Stability of PBNs with Respect to Noisy Domains

Our method to achieve stability of PBNs with respect to changes of the topological space K even under perturbations that change its topology, is to consider K embedded in a larger space X in which K and its noisy version are similar with respect to some metric.

Next we substitute the set K with an appropriate function f_K defined on X, so that the perturbation of the set K becomes a perturbation of the function f_K. As a consequence, instead of studying the shape of K as seen through a measuring function $\boldsymbol{\varphi}_{|K} : K \to \mathbb{R}^k$, we study a new measuring function $\boldsymbol{\Phi} : X \to \mathbb{R}^{k+1}$, with $\boldsymbol{\Phi} = (f_K, \boldsymbol{\varphi})$. PBNs of $\boldsymbol{\Phi}$ can be compared using the multidimensional matching distance in a stable way, as a consequence of stability with respect to functions perturbations. The key issue here is that we can prove that the PBNs of $\boldsymbol{\Phi}$ are still descriptors of the shape of K.

The proofs of the results presented in this section can be found in [15].

3.1 Comparison of Sets

A frequently used dissimilarity measure in classical set theory is the *Hausdorff distance*. If K_1, K_2 are non-empty compact subsets of $\mathbb{R}^n$, the Hausdorff distance can be defined by

$$\delta_H(K_1, K_2) = \max\{\max_{x \in K_2} d_{K_1}(x), \max_{y \in K_1} d_{K_2}(y)\},$$

where d_{K_1} and d_{K_2} denote the distance functions from K_1 and K_2, respectively.

The Hausdorff distance plays an important role in object recognition because it is quite resistant to small position errors such as those that may occur with feature extraction methods, but it is sensitive to outliers.

The symmetric difference pseudo-metric overcomes the problem of outliers. Denoting by μ the Lebesgue measure on $\mathbb{R}^n$, the *symmetric difference pseudo-metric* is defined between two measurable sets A, B with finite measure by

$$d_\triangle(A, B) = \mu(A \triangle B)$$

where $A \triangle B = (A \cup B) \setminus (A \cap B)$.

3.2 Stability with Respect to Hausdorff Distance

In order to achieve stability under set perturbations that are measured by the Hausdorff distance, we can take the function f_K equal to the function distance from K as the following result shows.

Theorem 1. *Let K_1, K_2 be non-empty closed subsets of a triangulable subspace X of $\mathbb{R}^n$. Let $d_{K_1}, d_{K_2} : X \to \mathbb{R}$ be their respective distance functions. Moreover, let $\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2 : X \to \mathbb{R}^k$ be vector-valued continuous functions. Then, defining $\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2 : X \to \mathbb{R}^{k+1}$ by $\boldsymbol{\Phi}_1 = (d_{K_1}, \boldsymbol{\varphi}_1)$ and $\boldsymbol{\Phi}_2 = (d_{K_2}, \boldsymbol{\varphi}_2)$, the following inequality holds:*

$$D_{match}\left(\beta_{\boldsymbol{\Phi}_1}, \beta_{\boldsymbol{\Phi}_2}\right) \le \max\left\{\delta_H(K_1, K_2), \|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|_\infty\right\}.$$

In plain words, Theorem 1 states that small changes in the domain and in the measuring function imply small changes in the PBNs, i.e. the shape descriptors.

The next result shows that the PBNs of $\boldsymbol{\Phi}$ still provide a shape descriptor for K as seen through $\boldsymbol{\varphi}_{|K}$.

Theorem 2. *Let K be a non-empty triangulable subset of a triangulable subspace X of $\mathbb{R}^n$. Moreover, let $\boldsymbol{\varphi} : X \to \mathbb{R}^k$ be a continuous function. Setting $\boldsymbol{\Phi} : X \to \mathbb{R}^{k+1}$, $\boldsymbol{\Phi} = (d_K, \boldsymbol{\varphi})$, for every $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^k$ with $\boldsymbol{u} \prec \boldsymbol{v}$, there exists a real number $\hat{b} > 0$ such that, for any $b \in \mathbb{R}$ with $0 < b \leq \hat{b}$, there exists a real number $\hat{a} = \hat{a}(b)$, with $0 < \hat{a} < b$, for which*

$$\beta_{\boldsymbol{\varphi}_{|K}}(\boldsymbol{u}, \boldsymbol{v}) = \beta_{\boldsymbol{\Phi}}\left((a, \boldsymbol{u}), (b, \boldsymbol{v})\right),$$

for every $a \in \mathbb{R}$ with $0 \leq a \leq \hat{a}$. In particular,

$$\beta_{\boldsymbol{\varphi}_{|K}}(\boldsymbol{u}, \boldsymbol{v}) = \lim_{b \to 0^+} \beta_{\boldsymbol{\Phi}}\left((0, \boldsymbol{u}), (b, \boldsymbol{v})\right).$$

In other words, Theorem 2 ensures that we can recover the PBNs of $\boldsymbol{\varphi}_{|K}$, i.e. a description of the shape of K as seen by $\boldsymbol{\varphi}$, from the PBNs of $\boldsymbol{\Phi}$, simply by passing to the limit.

3.3 Stability with Respect to the Symmetric Difference Pseudo-Distance

In order to achieve stability under set perturbations that are measured by the symmetric difference pseudo-distance, we can use as function f_K a function obtained convolving the characteristic function of K with that of a ball. More precisely, let $\lambda_K^\epsilon : \mathbb{R}^n \to \mathbb{R}$, with $\epsilon \in \mathbb{R}$, $\epsilon > 0$, be defined as

$$\lambda_K^\epsilon(x) = \mu(B_\epsilon)^{-1} \cdot \int_{y \in B_\epsilon(x)} \chi_K(y) \, \mathrm{d}y$$

where $B_\epsilon(x)$ denotes the n-ball centered at x with radius ϵ, $B_\epsilon = B_\epsilon(0)$, and χ_K denotes the characteristic function of K. In this case we have the following stability result.

Theorem 3. *Let K_1, K_2 be non-empty closed subsets of a triangulable subspace X of $\mathbb{R}^n$. Moreover, let $\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2 : X \to \mathbb{R}^k$ be vector-valued continuous functions. Then, defining $\boldsymbol{\Psi}_1^\epsilon, \boldsymbol{\Psi}_2^\epsilon : X \to \mathbb{R}^{k+1}$ by $\boldsymbol{\Psi}_1^\epsilon = (-\lambda_{K_1}^\epsilon, \boldsymbol{\varphi}_1)$ and $\boldsymbol{\Psi}_2^\epsilon = (-\lambda_{K_2}^\epsilon, \boldsymbol{\varphi}_2)$, the following inequality holds: $D\left(\beta_{\boldsymbol{\Psi}_1^\epsilon}, \beta_{\boldsymbol{\Psi}_2^\epsilon}\right) \leq \max\left\{\frac{d_\triangle(K_1,K_2)}{\mu(B_\epsilon)}, \|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|_\infty\right\}$.*

4 Experimental Results

In order to demonstrate the effectiveness of the approach presented here, we tested the method on the Kimia data set of 99 shapes [16], which are shown in Table 1. The dataset is classified in nine categories with 11 shapes in each category.

Each of the shapes has been corrupted by adding *salt & pepper* noise to a neighborhood of the set of its black pixels, as shown for some instances in Figure 2(Top). Salt & pepper noise is a form of noise typically seen on images, usually caused by errors in the data transmission. It appears as randomly

Table 1. Some instances from the database of 99 shapes with 9 categories and 11 shapes in each category used in our experiments. The complete database can be found in [16].

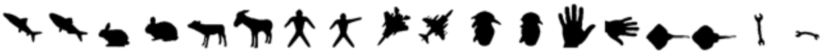
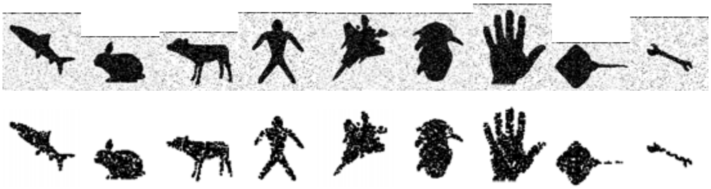


Table 2. Top: Shapes with salt & pepper noise. Bottom: The same shapes after morphological opening.



occurring white and black pixels, the percentage of pixels which are corrupted quantifying the noise. For each image, the set of black pixel of the image obtained by adding salt & pepper noise as in Figure 2(Top) is close to the set of black pixels of the original image with respect to the symmetric difference distance.

Salt & pepper noise can partially be removed by applying a morphological opening, thus obtaining shapes such as those in Figure 2(Bottom). The set of black pixels in the images so obtained is close to the set of black pixels of the original image with respect to the Hausdorff distance.

In both cases the topology of the set of black pixels in the noisy images is very different from that of the original images.

Three retrieval tests from the Kimia dataset of Table 1 were performed.

In order to provide a point of reference, the first retrieval test was performed without noise by matching each shape in the Kimia-99 dataset against every other shape in the database.

In the second retrieval test we used as models to be compared with all the shapes of the Kimia-99 database, the 99 images obtained by adding the salt & pepper noise and performing the morphological opening.

Finally, we compared the images corrupted by the salt & pepper noise with all the original images.

In all cases, ideal result would be that the 11 closest matches (including the queried model itself) all be of the same category as the query shape. The actual results we obtained are reported in Table 3. For each experiment, a string of 11 numbers describes the performance rate, the nth number corresponding to the rate at which the nth nearest match was in the same category as the model. This performance test has been applied to retrieval experiments from the Kimia-99 database by several authors testing their methods (see, e.g., [2,13,16,18,20]). However, our results are not directly comparable with theirs since we aim at a method tolerant under noise that modifies the shape topology.

Table 3. The retrieval rates of our method for the Kimia-99 database

Experiment	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th
without noise	99	95	91	88	85	82	80	76	63	53	40
with noise after opening	99	95	88	82	81	75	71	69	60	42	39
with noise without opening	99	91	87	78	76	71	69	62	57	45	38

We now describe how we obtained the results of Table 3. In each case we have used only the persistence diagrams of zeroth homology (a.k.a. size functions). Black pixels of each image represent the compact set K under study, respectively, whereas the black and white pixels of the bounding box constitute the ambient set X. A graph structure based on the local 8-neighbors adjacency relations of the digital points is used in order to topologize the images.

In the first experiment, without noise, for each shape we computed three persistence diagrams corresponding to the functions $\varphi_0, \varphi_1, \varphi_2 : X \to \mathbb{R}$ restricted to the set of black pixels K, where φ_0 is equal to minus the distance from the centroid of K, and φ_1, φ_2 are equal to minus the distance from the first and second axis of inertia of K, respectively.

In the second experiment, the query shapes were corrupted by noise and partially cleaned by the morphological opening. For each shape we computed 72 persistence diagrams: for each 2-dimensional function $\boldsymbol{\Phi}_0 = (d_K, \varphi_0) : X \to \mathbb{R}^2$, $\boldsymbol{\Phi}_1 = (d_K, \varphi_1) : X \to \mathbb{R}^2$, $\boldsymbol{\Phi}_2 = (d_K, \varphi_2) : X \to \mathbb{R}^2$, where $\varphi_0, \varphi_1, \varphi_2$ are as before and d_K is the distance from K, we obtain 24 persistence diagrams by considering the restriction of the associated Betti numbers to the planes of the foliation corresponding to the parameters $\boldsymbol{b} = (b, -b)$ with $b = 10, 13, 16$ and $\boldsymbol{l} = (\cos\theta, \sin\theta)$ with $\theta = 10°, 20°, \ldots, 80°$. The rationale behind these choices for b is that they ensure cooperation of d_K and φ_i in the function $F^{\varphi_i}_{(\boldsymbol{l},\boldsymbol{b})}$.

In the third experiment, the query shapes were corrupted by noise and no preprocessing was performed. For each shape we considered three 2-dimensional functions: $\boldsymbol{\Psi}_0 = (-\lambda^{\epsilon}_K, \varphi_0) : X \to \mathbb{R}^2$, $\boldsymbol{\Psi}_1 = (-\lambda^{\epsilon}_K, \varphi_1) : X \to \mathbb{R}^2$, $\boldsymbol{\Psi}_2 = (-\lambda^{\epsilon}_K, \varphi_2) : X \to \mathbb{R}^2$, where $\varphi_0, \varphi_1, \varphi_2$ are as before and λ^{ϵ}_K is the convolution of the characteristic function of K with that of the square of radius $\epsilon = 10$. By taking the restriction of the associated PBNs to the planes of the foliation corresponding to the parameters $\boldsymbol{b} = (b, -b)$ with $b = 3, 5, 7, 9$ and $\boldsymbol{l} = (\cos\theta, \sin\theta)$ with $\theta = 10°, 20°, \ldots, 80°$, for each shape we obtain 96 persistence diagrams. The motivation for these choices for b is the same as before.

In all three experiments persistence diagrams associated with each function (i.e. $\varphi_0, \varphi_1, \varphi_2$ in the first experiment, $\boldsymbol{\Phi}_0, \boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2$ in the second one, and $\boldsymbol{\Psi}_0, \boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2$ in the third one) were compared using the Hausdorff distance, as a lower bound of the matching distance to speed up computations. Next, these distances were normalized with mean equal to 0 and standard deviation equal to 1 so to obtain comparable values for different functions. Finally, as a dissimilarity measure between two shape, we took the sum of the normalized Hausdorff distances.

References

1. Belongie, S., Malik, J., Puzicha, J.: Shape Matching and Object Recognition Using Shape Contexts. IEEE Trans. Pattern Anal. Mach. Intell. 24(4), 509–522 (2002)
2. Bernier, T., Landry, J.A.: New method for representing and matching shapes of natural objects. Pattern Recognition 36(8), 1711–1723 (2003)
3. Biasotti, S., Cerri, A., Frosini, P., Giorgi, D., Landi, C.: Multidimensional size functions for shape comparison. J. Math. Imaging Vision 32(2), 161–179 (2008)
4. Biasotti, S., De Floriani, L., Falcidieno, B., Frosini, P., Giorgi, D., Landi, C., Papaleo, L., Spagnuolo, M.: Describing shapes by geometrical-topological properties of real functions. ACM Comput. Surv. 40(4) (2008)
5. Biswas, S., Aggarwal, G., Chellappa, R.: An Efficient and Robust Algorithm for Shape Indexing and Retrieval. IEEE Transactions on Multimedia 12(5), 372–385 (2010)
6. Cagliari, F., Di Fabio, B., Ferri, M.: One-dimensional reduction of multidimensional persistent homology. Proc. Amer. Math. Soc. 138(8), 3003–3017 (2010)
7. Cagliari, F., Landi, C.: Finiteness of rank invariants of multidimensional persistent homology groups. Applied Mathematics Letters 24, 516–518 (2011)
8. Carlsson, G., Zomorodian, A.: The theory of multidimensional persistence. Discrete & Computational Geometry 42(1), 71–93 (2009)
9. Cerri, A., Di Fabio, B., Ferri, M., Frosini, P., Landi, C.: Multidimensional persistent homology is stable. Technical Report, Università di Bologna, http://www.amsacta.cib.unibo.it/2603/
10. Cohen-Steiner, D., Edelsbrunner, H., Harer, J.: Stability of Persistence Diagrams. Discrete Comput. Geom. 37(1), 103–120 (2007)
11. Cohen-Steiner, D., Edelsbrunner, H., Harer, J., Morozov, D.: Persistent homology for kernels, images, and cokernels. In: SODA 2009: Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1011–1020 (2009)
12. d'Amico, M., Frosini, P., Landi, C.: Natural pseudo-distance and optimal matching between reduced size functions. Acta. Appl. Math. 109, 527–554 (2010)
13. Ebrahim, Y., Ahmed, M., Chau, S.-C., Abdelsalam, W.: A Template-Based Shape Representation Technique. In: Campilho, A., Kamel, M.S. (eds.) ICIAR 2008. LNCS, vol. 5112, pp. 497–506. Springer, Heidelberg (2008)
14. Edelsbrunner, H., Letscher, D., Zomorodian, A.: Topological persistence and simplification. Discrete & Computational Geometry 28(4), 511–533 (2002)
15. Frosini, P., Landi, C.: Stability of multidimensional persistent homology with respect to domain perturbations, arXiv:1001.1078v2 (2010)
16. Sebastian, T.B., Klein, P.N., Kimia, B.B.: Recognition of shapes by editing shock graphs. In: ICCV 2001, vol. 1, pp. 755–762 (2001)
17. Siddiqi, K., Shokoufandeh, A., Dickinson, S.J., Zucker, S.W.: Shock graphs and shape matching. Int. J. Comput. Vis. 35(1), 13–32 (1999)
18. Tu, Z., Yuille, A.L.: Shape matching and recognition – using generative models and informative features. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3023, pp. 195–209. Springer, Heidelberg (2004)
19. Verri, A., Uras, C., Frosini, P., Ferri, M.: On the use of size functions for shape analysis. Biol. Cybern. 70, 99–107 (1993)
20. Zhou, R., Zhang, L.: Shape retrieval using pyramid matching with orientation features. In: IEEE International Confenrence on Intelligent Computing and Intelligent Systems, vol. 4, pp. 431–434 (2009)

A Spanning Tree-Based Human Activity Prediction System Using Life Logs from Depth Silhouette-Based Human Activity Recognition

Md. Zia Uddin[1], Kyung Min Byun[2], Min Hyoung Cho[2], Soo Yeol Lee[2],
Gon Khang[2], and Tae-Seong Kim[2]

[1]Department of Electronic Engineering,
Inha University, 253 Yonghyun-dong, Nam-gu, Incheon, 402-751, Republic of Korea
[2]Department of Biomedical Engineering,
Kyung Hee University, Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do, 446-701,
Republic of Korea
ziauddin@inha.ac.kr,
{kmbyun,mhcho,sylee01,gkhang,tskim}@khu.ac.kr

Abstract. In this work, we propose a Human Activity Prediction (HAP) system using activity sequence spanning trees constructed from a life-log created by a video sensor-based daily Human Activity Recognition (HAR) system using time-sequential Independent Component (IC)-based depth silhouette features with Hidden Markov Models (HMMs). In the daily HAR system, the IC features are extracted from the collection of the depth silhouettes containing various daily human activities such as walking, sitting, lying, cooking, eating etc. Using these features, HMMs are used to model the time sequential features and recognize various human activities. The depth silhouette-based human activity recognition system is used to recognize daily human activities automatically in real time, which creates a life-log of daily activity events. In this work, we propose a method for human activity prediction using fixed-length activity sequence spanning trees based on the life-log. Utilizing the consecutive activities recorded in an activity sequence database (i.e. life-log) for a specific period of time of each day over a period such as a month, the fixed-length spanning trees can be constructed for the sequences starting with each activity where the leaf nodes contain the frequency of the fixed-length consecutive activity sequences. Once the trees are constructed, to predict an activity after a sequence of activities, we traverse the spanning trees until a path up to the previous node of the leaf nodes is matched with the testing pattern. Finally, we can predict the next activity based on the highest frequency of the leaf nodes along the matched path. The prediction experiments over the computer simulated data which is based on the daily logs show satisfactory results. Our video sensor-based human activity recognition and prediction systems can be utilized for practical applications such as smart and proactive healthcare.

Keywords: PCA, ICA, LDA, HMM, Spanning Tree, Human Activity Recognition, Human Activity Prediction.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 302–309, 2011.

1 Introduction

In the recent years, Human Activity Recognition (HAR) is getting considerable attention among the researchers of Human Computer Interaction (HCI) [1]-[3]. So far, in the video-based HAR, binary silhouettes from RGB activity videos are the most commonly employed from which useful features are derived [4], [5]. In [4] and [5], Uddin et al proposed local features from the binary silhouettes via Independent Component Analysis (ICA) to recognize different human activities.

However, binary silhouettes are not efficient enough to describe human body properly in the activity videos due to its two-level flat pixel intensity distribution although they have been extensively utilized. Depth values can represent the human body better than the binary representation by differentiating the body parts by means of different intensity values [6], [7]. With a depth silhouette-based real-time HAR system, one can recognize various daily human activities such as walking, eating, lying, sitting, cooking etc. Thus, a HAR system acts as a life-log agent that logs what activities a human subject performs everyday in a specific time. There have been a few attempts to create a log a person's life in different research areas [8], [9]. Using the human activity log information, many researchers have tried to predict human activities [10], [11]. In [10], the authors proposed a behavior prediction system to support daily lives where the behaviors in a daily life were recorded with some embedded sensors, and the prediction system learned the characteristic patterns that would be followed by the behaviors to be predicted. The prediction system observes daily behaviors with sensors and outputs the prediction of future behaviors based on some rules. For their experiments, the authors applied 1,250 rules for prediction. In another work [11], the authors focused on the prediction of the progression of a particular activity on the basis of a 24-hour period to detect an unexpected event which could indicate a change in a health condition. In general, humans habitually repeat the same kind of sequential patterns every day and hence, the contiguous activity sequences can be focused for efficient Human Activity Prediction (HAP).

In this work, we propose an activity sequence spanning tree-based HAP system using life-log created through a real-time depth silhouette-based HAR system. Then, we build fixed-length activity spanning trees containing the frequency of the contiguous activity sequences. Every node in the tree represents an activity. After recognizing some contiguous activities automatically by a HAR system in real-time, we predict the next activity going to be performed by means of the activity spanning tree information. Thus, we traverse the tree for matching the contiguous activity sequence for prediction. Along with that matched path in the tree, there could be more than one leaf nodes. Then, considering the leaf node with the highest frequency in the matched path in a tree, we predict the next activity going to be performed. With the depth-based HAR system, we have performed validation with the simulated data to test our prediction system.

The remaining sections of our paper are structured as follows. Section 2 describes the methodology of HAR system. Section 3 shows the basic steps to create a life-log using our HAR system. Section 4 represents the spanning tree-based HAP. Section 5 deals with the experiments and results. At last, Section 6 provides the concluding remarks.

2 Depth Silhouette-Based Human Activity Recognition

In the depth silhouette-based HAR, the process starts with depth silhouette extraction from the time sequential activity video images. The RGB and depth images of five home activities including eating, lying, sitting, cooking, and walking are acquired by ZCAM^{TM} (3DV Systems Ltd) [12]. Fig. 1 demonstrates HMM-based HAR procedure as well as normalized basis images in a gray scale after applying PCA, ICA, and LDA on the IC features over the depth silhouettes of the five activities: namely walking, lying, sitting, cooking, and eating. For HAR experiments, our depth silhouette-based

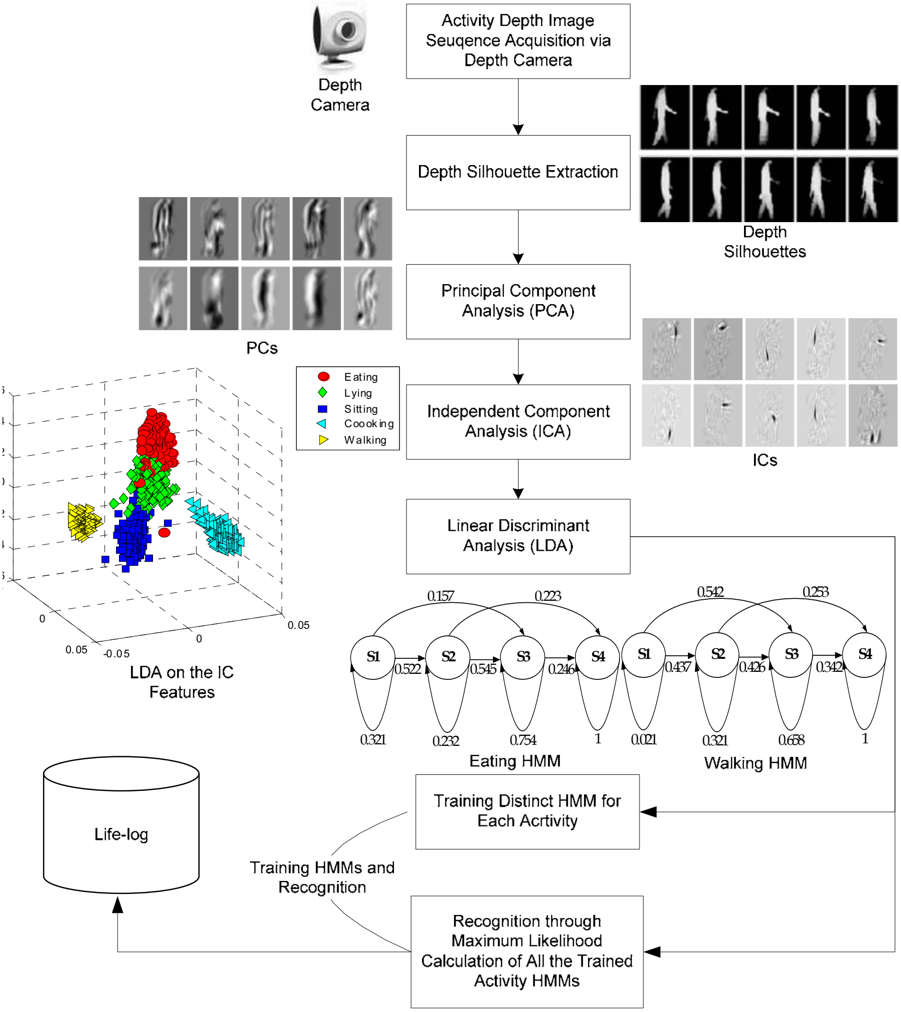

Depth Camera
Activity Depth Image Seuqence Acquisition via Depth Camera
Depth Silhouette Extraction
Depth Silhouettes
PCs
Principal Component Analysis (PCA)
Independent Component Analysis (ICA)
ICs
Eating
Lying
Sitting
Coooking
Walking
0.05
0
-0.05
0
0.05
LDA on the IC Features
Linear Discriminant Analysis (LDA)
0.157
0.223
S1
0.522
S2
0.545
S3
0.246
S4
0.321
0.232
0.754
1
Eating HMM
0.542
0.253
S1
0.437
S2
0.426
S3
0.342
S4
0.021
0.321
0.658
1
Walking HMM
Training Distinct HMM for Each Acrtivity
Life-log
Training HMMs and Recognition
Recognition through Maximum Likelihood Calculation of All the Trained Activity HMMs

Fig. 1. Basic steps for depth silhouette-based real-time HAR to create an activity life-log

activity database was built for five activities (i.e., walking, lying, sitting, cooking, and eating) where each clip contained variable length consecutive frames. In order to train and test each activity HMM, we applied 15 and 40 image sequences respectively. We applied LDA on the ICA features of binary and depth silhouettes for the experiments with HMM for training and recognition. The depth silhouette–based HAR approach achieved the mean recognition rate of 96.50%, which is superior to binary silhouette-based approach that achieved mean rate of 88.50%.

3 Generation of a Life-Log from Depth Silhouette-Based HAR System

The depth silhouette-based HAR system was applied in real-time to recognize the five human activities automatically and saved the activity events with time stamps as a life-log. Once a life-log is available containing the activities of each day, human activity prediction can be done based on contiguous sequential activity patterns. Table 1 shows an sample database of daily logs of activities where the letters represent the activities such as W for walking, L for lying, S for sitting, E for eating, and C for cooking.

Although we can generate a short daily log of activities with the real-time HAR system, to test the HAP system, one needs an extensive database of daily logs. Thus, an activity life-log database is created by means of a program that randomly generates the activity sequences similar to the events in Table 1. Two random numbers are generated for the dataset where the first one is to generate the random activity among five activities and the second one to determine the number of repetitions of the most recent activity. In the next section, we are going to discuss an idea to predict human activity using spanning tree containing activity sequences with their frequencies.

4 Spanning Tree-Based Human Activity Prediction

To start the prediction process, we require a life-log of human activities for a specific period of every day for one month or two months. Each day's time-line is divided by a fixed number of seconds (such as 10 sec) and tagged with what activities are performed in that period. Based on the tagged information, we have built fixed-length spanning trees containing activity sequences where each node represents an activity and the leaf nodes are used to predict using the upper level nodes. Fig. 2 shows the basic processes to create activity sequence spanning trees to use them for HAP.

Table 1. A sample activity database for six days

Day	Activity Sequence
1	CCCCSSSSCCCCSSS
2	SSCCCCCLLLLLLLL
3	SSSLLLLLLCCCCWW
4	EEEEEWWWCCCCLLL
5	CCCCCWWWWWCCCCC
6	WWWWWWWWWWWWSSS

Now, we run a fixed-length window through each row of the database and build fixed-length spanning trees where the leaf node stores the frequency of fixed-length patterns. Fig. 3(a) shows a five-length sliding window approach to scan the activity database shown in Table 1. Figs. 3(b) to (f) show the W-tree, L-tree, S-tree, E-tree, and C-tree respectively. Table 2 summarizes our algorithm where it is divided in two parts. The first part contributes to build fixed-length activity spanning trees which is used in the second part for prediction. The second part shows how to predict F^{th} activity using a fixed-length activity sequence containing the length of (F-1).

Table 2. Proposed algorithm for activity spanning tree construction and human activity prediction

```
Algorithm
PART1: Constructing Spanning Trees
Step: Extract fixed-length activity sequences from database and construct spanning trees
//Access all database sequences
1. for(i=0; i< N; i++)
//Extract fixed-length subsequences
2. FS = Extract_Fixed_Length_Subseq(F, $S_i$);
// Construct spanning trees with the fixed-length pattern and it's frequency in leaf node.
3. SP_Tree= Construct_Spanning_Trees(F);
PART2: Prediction Using the Constructed Spanning Trees
Step : Search trees with an activity sequence of length (F-1) $S_T$ and try to predict the $F^{th}$ activity.
//Extract matched sequence in the tree up to level four.
6. Matched_Seq= Search_in_Spanning_Tree(SP_Tree, $S_T$);
//Observe the frequencies of the leaf nodes along with the matched path in the tree. The leaf node //with highest frequency can be the next activity. If the leaf nodes' frequencies are almost same or //marginally greater or lower than no prediction can work and hence return NULL.
7. Activity_F = CheckFreqLeaf(SP_tree, Matched_Seq);
```

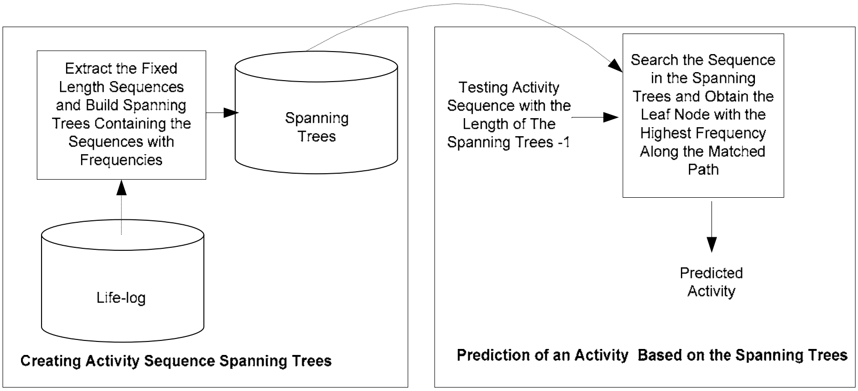
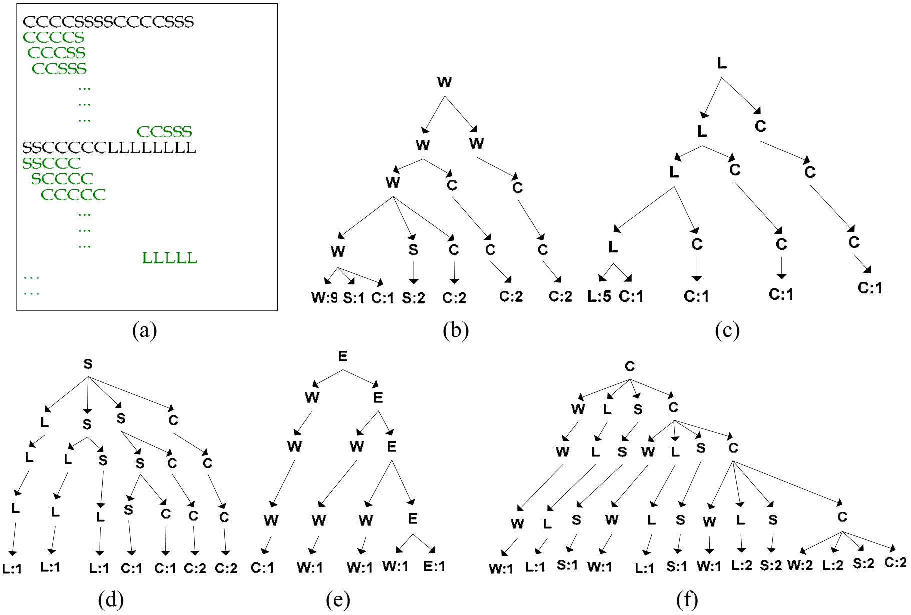

Extract the Fixed Length Sequences and Build Spanning Trees Containing the Sequences with Frequencies
Spanning Trees
Life-log
Creating Activity Sequence Spanning Trees
Testing Activity Sequence with the Length of The Spanning Trees -1
Search the Sequence in the Spanning Trees and Obtain the Leaf Node with the Highest Frequency Along the Matched Path
Predicted Activity
Prediction of an Activity Based on the Spanning Trees

Fig. 2. Processes to create activity sequence spanning trees for human activity prediction
]

As our tree is of five-length, we recognize activities in consecutive four time slots (i.e., 40 seconds where 10 seconds for each slot) and try to predict the fifth activity to be performed in the next 10 seconds using the activity spanning trees. Thus, we start matching the four-length consecutive tree pattern in the trees and if there is any match, we can find the leaf nodes in that path and their frequencies.



(a) (b) (c)

(d) (e) (f)

Fig. 3. Five-length (a) Scanning (b) W-, (c) L-, (d) S-, (e) E-, and (f) C-tree

If there is only one node, there will be a good possibility of that activity for being next activity. If there are many, we see that which one contains the highest frequency and then predict that activity as the next activity. Besides, we apply here a threshold on the highest frequency i.e., the highest frequency should be greater than the threshold percentage of the total frequency of all the leaf nodes along the matched path. For instance, we want to recognize four consecutive activities as WWWW and now we would like to predict the next activity to be performed. We traverse the trees for the pattern and get a match in W-tree. However in that path, there are three leaf nodes as W with frequency of nine, S with frequency of 1 and C with frequency of 1. As frequency of W is much higher than that of others, we predict that the next activity should be W (i.e., walking). Another instance is CCCC where we can see that in that path, all the leaf nodes have the same frequency. So, we cannot predict the next activity here and hence, we continue recognition for the next time slot without prediction. Later on, we try to predict again using the spanning trees.

5 Experiments and Results

We obtained a total number of 60 activity sequences for 60 days in the database created through simulation where each sequence was 2,048 in length and consists of five distinct items (i.e., W, L, S, C, and E) to represent different activities. Using these 60 days data sequences we created the five-length spanning trees for each activity. Later on, to test our prediction approach, we created another datasets which consists of 6,210 test patterns of five-length to predict the 5th activity. For every test pattern, we tried to match a path of four-length in the spanning trees and based on the frequencies of the leaf nodes in the matched path, the 5th activity was predicted with which the 5th activity of the test pattern was compared to verify the prediction. Finally, we tested our approach to test all the testing patterns and obtained the mean prediction rate of 91.31% successfully. Besides, during prediction, we applied a threshold to check the frequency of the leaf node containing the maximum frequency was greater than or equal to 80% of total frequency of the leaf nodes along the matched path or not. When any leaf node in that path satisfied the condition, we continued prediction otherwise no predication for that sequence and we proceeded to the next testing sequence for prediction.

6 Conclusion

In this work, a novel fixed-length activity spanning tree-based activity prediction approach, based on the life-log created through human activity recognition using depth silhouette features with HMM, has been proposed. Utilizing our HAP approach, we obtained 91.31% prediction results using the computer simulated activity sequence datasets. Our HAP system still requires real tests under real environments, but could be used for human activity recognition and prediction at an environment like smart homes.

Acknowledgement. This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2011-(C1090-1121-0003).

References

1. Robertson, N., Reid, I.: A General Method for Human Activity Recognition in Video. Computer Vision and Image Understanding 104(2), 232–248 (2006)
2. Kang, H., Lee, C.W., Jung, K.: Recognition-based gesture spotting in video games. Pattern Recognition Letters 25, 1701–1714 (2004)
3. Chen, F.S., Fu, C.M., Huang, C.L.: Hand gesture recognition using a real-time tracking method and Hidden Markov Models. Image and Vision Computing 21, 745–758 (2005)
4. Uddin, M.Z., Lee, J.J., Kim, T.-S.: Independent Component Feature-based Human Activity Recognition via Linear Discriminant Analysis and Hidden Markov Model. In: Proceedings of 30th Annual International IEEE EMBS Conference, pp. 5168–5171 (2008)

5. Uddin, M.Z., Lee, J.J., Kim, T.-S.: Shape-Based Human Activity Recognition Using Independent Component Analysis and Hidden Markov Model. In: Proceedings of the 21st International Conference on Industrial, Engineering, and Other Applications of Applied Intelligent Systems: New Frontiers in Applied Artificial Intelligence, pp. 245–254 (2008)
6. Uddin, M.Z., Lee, J.J., Kim, T.-S.: Human Activity Recognition Using Independent Component Features from Depth Images. In: Proceedings of the 5th International Conference on Ubiquitous Healthcare, pp. 181–183 (2008)
7. Munoz-Salinas, R., Medina-Carnicer, R., Madrid-Cuevas, F.J., Carmona-Poyato, A.: Depth silhouettes for gesture recognition. Pattern Recognition Letters 29, 319–329 (2008)
8. Hori, T., Aizawa, K.: Sigmultimedia, Context-based video retrieval system for the life-log applications. In: Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval, pp. 31–38 (2003)
9. Mann, S.: wearcam (the wearable camera): personal imaging system for long-term use in wearable tether less computer-mediated reality and personal photo/video graphic memory prosthesis. In: Proceedings of IEEE ISWC 1998, pp. 124–131 (1998)
10. Mori, T., Takada, A., Noguchi, H., Harada, T., Sato, T.: Behavior Prediction Based on Daily-Life Record Database in Distributed Sensing Space. In: Proceedings of the International Conference on Intelligent Robots and Systems, pp. 1703–1709 (2005)
11. Cook, D.J., Youngblood, M., Heierman III, E.O., Gopalratnam, K., Rao, S., Litvin, A., Khawaja, F.: MavHome: an agent-based smart home. In: Proceedings of the First IEEE International Conference on Pervasive Computing and Communications, pp. 521–524 (2003)
12. Iddan, G.J., Yahav, G.: 3D imaging in the studio (and elsewhere...). In: Proceedings of SPIE, vol. 4298, pp. 48–55 (2001)

Characterizing Obstacle-Avoiding Paths Using Cohomology Theory

Paweł Dłotko[1], Walter G. Kropatsch[2], and Hubert Wagner[1]

[1] Institute of Computer Science, Jagiellonian University, Poland
{hubert.wagner,dlotko}@ii.uj.edu.pl
[2] Pattern Recognition and Image Processing Group,
Vienna University of Technology, Austria
krw@prip.tuwien.ac.at

Abstract. In this paper, we investigate the problem of analyzing the shape of obstacle-avoiding paths in a space. Given a d-dimensional space with holes, representing obstacles, we ask if certain paths are equivalent, informally if one path can be continuously deformed into another, within this space. Algebraic topology is used to distinguish between topologically different paths. A compact yet complete *signature* of a path is constructed, based on cohomology theory. Possible applications include assisted living, residential, security and environmental monitoring. Numerical results will be presented in the final version of this paper.

Keywords: obstacle-avoidance, cohomology generators, trajectory planning problem.

1 Introduction

In the recent years, there has been growing interest in topics such as assisted living, residential, security and environmental monitoring [1,2]. This is closely related to the area of remote sensing, which aims at delivering a description of the chosen aspects of the sensed environment by aggregating information from an array of sensors.

The information gathered by individual sensors ranges from visual data (Visual Sensor Networks [3]) to the presence of smoke in the air. Visual Sensor Networks are the most closely related to the computer vision field. In this paper we treat the sensors in an abstract way, therefore the method should be applicable in a number of settings.

One important question that arises is how to arrange such sensors. In [1], which largely inspired us to write this paper, straight laser beams are used as sensors. Prompted by some of the questions posed in the summary of that paper, we consider the following questions. How does this scenario generalize to sensors of different shapes? Can we generalize these concepts to higher dimensions (the original considerations were done in 2D)?

As a simple example, consider a network of paid highways. Since exact tracking the movement of each vehicle is prohibitively expansive, simplified measurements

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 310–317, 2011.

have to be performed. Gates serve as sensors, enabling us to roughly estimate the movement of the vehicle. While we fail to capture the precise *geometry* of the path of the vehicle, we are able to capture what we consider the *topology*.

This is closely related to the recent concept of *minimal sensing*, where sensors are very limited in their capabilities. In such a setting, sensors are typically unable to capture the actual geometry of the space. See [2] and references therein, to see how this problem was tackled, often using algebraic topology.

While the above example is trivial and can be described with basic graph algorithms, the situation is much more interesting (and challenging) in higher dimensions. Since our approach is based on algebraic topology, especially cohomology theory, it is dimension-independent.

Our additional aim is to to expose cohomology theory to the CAIP community. We believe that the mathematical robustness and intuitivity make it an interesting tool, which can be applied more generally.

The paper is structured as follows: In Section 2 a rigorous formulation of the considered problem is presented. In Section 3 the complexes used in this paper are discussed. In Section 4 an intuitive introduction to homology and cohomology theory is given. In Section 5 the main result of this paper is stated. In Section 6 an algorithm to compute signature of a given path is presented. Finally in Section 7 the conclusions are drawn.

2 Problem Formulation

We analyze movement, from point S to T, of a number of agents in a known space. Simply put, we ask how to place sensors, so that we are able to describe the *topology* of each path, based only on how it intersects these sensors. We encode these intersections as a *signature*, which is sufficient to discriminate between paths having different topology (more precisely: homology). We will prove that the sensors need to be placed in the support of *cohomology generators*.

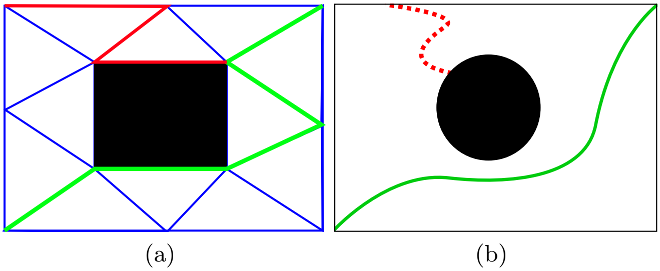
The problem of analyzing paths of moving agents in a 2-dimensional space, in the presence of obstacles and linear (beam) sensors was introduced in [1]. We present a variation for a d-dimensional, orientable space, where "sensors" are represented by certain $(d-1)$-dimensional hypersurfaces (possibly with self-intersections). For the $2-$dimensional case the difference is that our sensors can have arbitrary shape and are allowed to intersect. While the idea of a sensor of arbitrary shape might seem contrived, imagine that such a sensor is actually composed of a number of small sensing units covering a given hypersurface.

3 Representing Spaces with Holes

In this section we present some theory related to computational topology, used later in the paper. For simplicity the concept of *simplicial complex* is used to represent the space. The definition of simplicial complex can be found in [4]. Imagine that a simplicial complex is a decomposition of the space into a set of simplices, that is vertices, edges, triangles etc. In general, $n-$simplex is a convex hull of $n+1$ points lying in general position. The number n is the *dimension* of a simplex S and is denoted by $dim(S)$. We assume that vertices of a simplicial complex are uniquely enumerated with integers, allowing to index each simplex with the set of its vertices. Each simplex in the simplicial complex has an orientation (this is discussed in details in [5]). In the implementation presented in [6], enumeration of vertices of complex $\mathcal{K}$ is used in orienting the edges and higher dimensional simplices. For instance every edge E is oriented from its higher vertex to lower vertex. From now on the orientation of all simplices in the complex is assumed to be fixed. A subset of simplices is chosen to represent the obstacles. During the computation of cohomology, the interior of obstacles is removed from the complex. Later by $\mathcal{K}$ we will denote the complex after this removal.

There are two vertices chosen in our complex, marked as S and T from Source and Target. An oriented path is the formal sum of edges joining those points with $+1, -1$ coefficients, which induce orientation.

The goal is to provide an efficient algorithm to describe and distinguish paths from S to T, which avoid all the obstacles[1]. An example of a 2−dimensional simplicial complex can be found in Figure 1(a).


(a)

(b)

Fig. 1. a) Simple example of a complex. Obstacles are marked with black, paths with green (solid). b) Graphical representation of complexes that we will use for clarity of images. Imagine that the complex is very finely subdivided, but paths and generators are still composed of edges of the complex, which is not displayed. Cohomology generator is depicted as the red (dotted) curve. In both cases point S is placed in the lower left, and point T in the upper right corner of the picture.

4 Cohomology Theory

In this section an intuitive exposition of homology and cohomology theory is given. For a full introduction consult [5]. Both homology and cohomology groups give a compact description of topology of a simplicial complex.

In homology theory one uses a concept of *chain*, being a formal sum of simplices with integer coefficients. A group of chains of dimension n is denoted by $C_n(\mathcal{K}) := \{\sum_{S\in\mathcal{K},dim(S)=n} \alpha_S S\}$. A boundary operator $\partial : C_n \to C_{n-1}$ is then introduced for a simplex $S = [v_0, \ldots, v_n]$:

[1] Note that the number of homologically different paths is unbounded for non-trivial cases.

$$\partial S = \sum_{i=0}^{n} (-1)^i [v_0, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n] \tag{1}$$

and extended linearly to $C_n(\mathcal{K})$. As an example, let us calculate the boundary of a full triangle: $\partial[0,1,2] = [1,2] - [0,2] + [0,1]$.

A group of n dimensional *cycles* $Z_n(\mathcal{K}) := \{c \in C_n(\mathcal{K}) \mid \partial c = 0\}$. In short, a cycle is a chain whose boundary vanishes. A group of $n-$ dimensional *boundaries* $B_n(\mathcal{K}) := \{c \in C_n(\mathcal{K}) | \exists d \in C_{n+1}(\mathcal{K}) | \partial d = c\}$. The idea behind cycles and boundaries is presented in Figure 2(a).

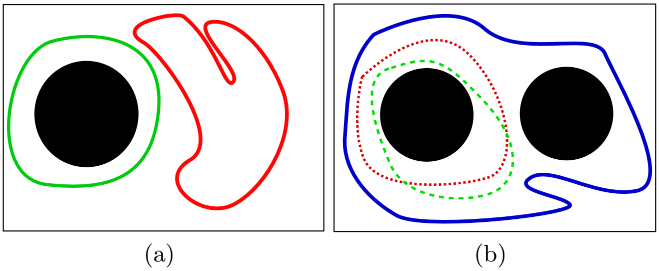

(a) (b)

Fig. 2. a) Right chain is a cycle and a boundary. Left cycle surrounds a hole, so it is not a boundary b) Red (dotted) and green (dashed) cycles are homologous. Blue (bold) is not homologous with any of them. Red (or green) and blue cycles constitute a homology basis.

It is straightforward to verify from Formula 1 that $\partial\partial = 0$. Therefore we have $B_n(\mathcal{K}) \subset Z_n(\mathcal{K})$ and we can define the *homology group* $H_n(\mathcal{K})$ as a classes of cycles which are not boundaries, namely $H_n(\mathcal{K}) := Z_n(\mathcal{K})/B_n(\mathcal{K})$. Two n-cycles c_1 and c_2 such that $c_1 - c_2 \in B_n(\mathcal{K})$ are said to be *homologous.* By homology generators we mean any representants of classes of cycles that generate $H_n(\mathcal{K})$. In absence of *torsions* the rank of homology group can be interpreted as number of holes in the considered space. Idea of homology groups is given in Figure 2(b).

In this paper we restrict ourselves to connected simplicial complexes $\mathcal{K}$ which are torsion-free in dimension one (i.e. after the obstacles are removed from the complex, the resulting complex is connected and torsion free). Torsions in homology mean that elements of a homology group have finite order (they generate a subgroup $\mathbb{Z}_p$ of homology group for $p \in \mathbb{Z}$ being the order of an element).

For a formal introduction to the cohomology theory consult [5], for an intuitive introduction consult [6]. Further in the paper we need a concept of *n-cochain* c^* being a map assigning any chain $c \in Z_n(\mathcal{K})$ a number[2] $\langle c^*, c\rangle \in \mathbb{Z}$. A group

[2] Operation $\langle c^*, c\rangle$ is called evaluation of a cocycle c^* on a cycle c. In order to compute $\langle c^*, c\rangle$, note that the set of maps $\{S^* | \langle S^*, K\rangle = \delta_{SK}$ for any $K \in \mathcal{K}\}_{S\in\mathcal{K}}$ constitutes a basis of $C^n(\mathcal{K})$. Therefore every cochain c^* is equal to $\sum_{S\in\mathcal{K}} \alpha_S S^*$ for *dim* $S = n$. Then for a chain $c = \sum_{S\in\mathcal{K}} \beta_S S$ we have $\langle c^*, c\rangle = \langle \sum_{S\in\mathcal{K}} \alpha_S S^*, \sum_{S\in\mathcal{K}} \beta_S S\rangle = \sum_{S\in\mathcal{K}} \alpha_S \beta_S$.

of $n-$cochains is denoted as $C^n(\mathcal{K})$. Dually to homology, a so-called *coboundary operator* $\delta : C^n(\mathcal{K}) \to C^{n+1}(\mathcal{K})$ is introduced. It is defined as $\langle \delta c^*, c\rangle = \langle c^*, \partial c\rangle$ for every $c^* \in C^{n-1}(\mathcal{K})$ and $c \in C_n(\mathcal{K})$. Again, cochain c^* is a *cocycle* if $\delta c^* = 0$. Cochain c^* is a *coboundary* if there exists a cochain $d^* \in C^{n-1}(\mathcal{K})$ such that $\delta d^* = c^*$. Cocycles are denoted as $Z^n(\mathcal{K})$, and coboundaries as $B^n(\mathcal{K})$. Finally, *cohomology group* is defined as the quotient $H^n(\mathcal{K}) := Z^n(\mathcal{K})/B^n(\mathcal{K})$.

It might appear that for torsion-free spaces all (co)homology computations could be performed with $\mathbb{Z}_p$ coefficients for $p \in \mathbb{Z}$, $p \geq 2$. This is not the case. Without going into details: we must use $\mathbb{Z}$ coefficients to handle the case of paths crossing certain cohomology generators np-times for $n \in \mathbb{Z}$.

For our purposes it is sufficient to consider cohomology group basis in dimension one. For torsion-free spaces, there is a straightforward correspondence between homology and cohmology group generators (see Theorem 4.8, [7]). Theorem 4.8 states that for any set of cycles representing homology generators $h_1, \ldots, h_n$ there exist dual cohomology generators $h^1, \ldots, h^n$ such that $\langle h^i, h_j\rangle = \delta_{ij}$. This theorem allows us to use the so-called "cutting analogy" to describe a cohomology basis. In fact, in the considered case the generator h^i, for $i \in \{1, \ldots, n\}$, can be seen as a fence that blocks any cycles in the class of h_i. This idea is illustrated in Figure 3(a). The concept of the presented "cut analogy" was developed in the so-called *Discrete Geometrical Approach* to Maxwell's equations [6].

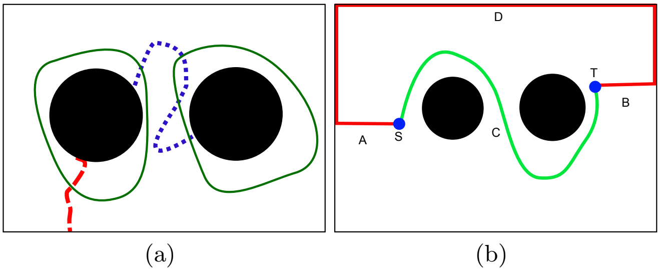


Fig. 3. a) The "cut analogy". When one cuts a complex along the red (dashed) cohomology generator, the left homology class vanishes. Cutting along blue (dotted) generator makes the right homology class vanish. b) Completion of chain c.

With the algorithm described in [6], we obtain cohomology generators (represented as a set of pairs (edge, integer)) of any simplicial complex. Note that cohomology generators are allowed to intersect. See the Borromean Rings phenomenon in [8] for an example of a 3-dimensional space, where it is impossible to find a non-intersecting cohomology basis.

5 Path Characterization Using Signatures

In this section a formal proof of the main result of the paper is provided. Suppose a simplicial complex $\mathcal{K}$ is given. As previously, we assume that $H_1(\mathcal{K})$ is torsion-free and $\mathcal{K}$ itself is connected. Let $h^1, \dots, h^n$ be cocycles representing first cohomology group generators of $\mathcal{K}$. Moreover, let $h_1, \dots, h_n$ be the homology generators dual to $h^1, \dots, h^n$ according to Theorem 4.8 in [7] (they are only needed for the proof). We fix $h^1, \dots, h^n$ and their dual $h_1, \dots, h_n$ for the rest of this section. Let $c \in C_1(\mathcal{K})$ be a path from S to T.

Definition 1. *For a path c the vector $S_c = [a_1, \dots, a_n]$ such that $a_i = \langle h^i, c \rangle$, for $i \in \{1, \dots, n\}$, is called a* signature *of c.*

In this section we show that paths having the same signature are homologous and, conversely, that paths having different signature are non-homologous. It is necessary to assume that all paths lead from S to T. A signature of a path provides an efficient way of distinguishing non-homologous paths and identifying homologous ones. Let us start with a lemma, the proof of which can be found in [7].

Lemma 1. *Let $c^* \in Z^1(\mathcal{K})$ be a cocycle and let $b \in B_1(\mathcal{K})$ be a boundary. Then $\langle c^*, b \rangle = 0$.*

Let us now define the *completion* of a chain. Let us take any chain A joining point S with the boundary of the complex $\mathcal{K}$, B joining point T with the boundary of a complex and D lying entirely on the boundary of $\mathcal{K}$ joining endpoints of chains A and B. With any path $c \in C_1(\mathcal{K})$ from S to T we can assign a cycle $c \cup A \cup B \cup D$. This cycle is called a *completion* of chain c (see Figure 3(b)).

Now we are ready to give the two main theorems of this paper.

Theorem 1. *Two homologous paths c_1 and c_2 have the same signature, $S_{c_1} = S_{c_2}$.*

Proof. Since c_1 and c_2 are homologous, there exists $b \in C_2(\mathcal{K})$ such that $\partial b = c_1 - c_2$. Therefore $c_1 = c_2 + \partial b$. From Lemma 1 we have, that $\langle h^i, c_1 \rangle = \langle h^i, c_2 + \partial b \rangle = \langle h^i, c_2 \rangle + \langle h^i, \partial b \rangle = \langle h^i, c_2 \rangle + 0 = \langle h^i, c_2 \rangle$ for every $i \in \{1, \dots, n\}$. Therefore $S_{c_1} = S_{c_2}$. □

Theorem 2. *Two non-homologous paths c_1 and c_2 have different signatures, $S_{c_1} \neq S_{c_2}$.*

Proof. Suppose by contrary that c_1 and c_2 are non-homologous and $S_{c_1} = S_{c_2}$. Therefore $d_1 = c_1 \cup A \cup B \cup D$ and $d_2 = c_2 \cup A \cup B \cup D$ are also non-homologous. But $h_1, \dots, h_n$ is a homology basis dual to cohomology basis $h^1, \dots, h^n$. Then we have $d_1 = \sum_{i=1}^n \alpha_i h_i + \partial e$ and $d_2 = \sum_{i=1}^n \beta_i h_i + \partial f$ for some $e, f \in C_2(\mathcal{K})$ and $\alpha_i, \beta_i \in \mathbb{Z}$ for $i \in \{1, \dots, n\}$. Since d_1 and d_2 are not homologous there exists an index $i \in \{1, \dots, n\}$ such that $\alpha_i \neq \beta_i$. But from the hypothesis we have $S_{c_1} = S_{c_2}$. It implies, that $S_{d_1} = S_{d_2}$. We have $\langle h^i, d_1 \rangle = \langle h^i, \sum_{i=1}^n \alpha_i h_i \rangle = \alpha_i$ and $\langle h^i, d_2 \rangle = \langle h^i, \sum_{i=1}^n \beta_i h_i \rangle = \beta_i$. Therefore from the hypothesis we have $\alpha_i = \beta_i$ for every $i \in \{1, \dots, n\}$, which gives a contradiction. □

6 Computing the Signature of a Path

In this section we present an algorithm which, for fixed cocycles $h^1, \ldots, h^n$, constituting a cohomology basis and a path c from A to B outputs S_c, the signature of c. We assume that simplicial complex is represented as a pointer-based data-structure as in [6]. Moreover, let each edge E of simplicial complex $\mathcal{K}$ be equipped with a vector v of n integers such that $v_E[i] = \langle h^i, E \rangle$ for every $i \in \{1, \ldots, n\}$. Let a path c be given as a vector of pointers to edges in $\mathcal{K}$.

It remains to resolve the subtlety of orientation of simplices versus an orientation of a path c. The path is oriented from point S to T. Let us define $o(c, E)$ in the following way: $o(c, E) := 1$ if orientation of c is the same as orientation of E and -1 otherwise. Now we list the algorithm. Also, see Figure 4 for a visual example. Note that this two-dimensional example is very simple and can be solved with basic tools, but our method works for general dimension.

Algorithm 1. Computing signature of a path

```
Input: path $c$, simplicial complex $\mathcal{K}$ with cohomology generators $h^1, \ldots, h^n$
Output: $s$ - signature of path $c$
Let $v$ be the vector encoding the intersections of $c$ with cohomology generators
Let $s$ be an $n$-tuple
for $i \in \{1, \ldots, n\}$ do
    $s[i] \leftarrow \sum_{E \in c} o(c, E) v_E[i]$
return $s$
```

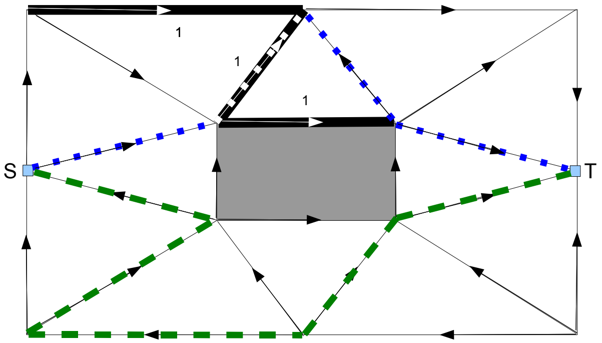


Fig. 4. We use the presented procedure to compute $s[1]$ for the blue (dotted) path. $v_E[1]$ is nonzero only for the edges in the support of the cohomology generator. Therefore $s[1] = 1$, as the orientation of this path is the same as the orientation of cohomology generator (bold black). As for the green (dashed) path $s[1] = 0$, since the path do not cross the cohomology generator. Blue (dotted) and green (dashed) paths are clearly non-homologous.

7 Conclusions

The ideas presented in this paper generalize the approach using "laser beams" presented in [1]. We use topological tools to distinguish between different obstacle-avoiding paths, based only on their intersections with selected *sensors*. The usage of algebraic topology enables us to use sensors of arbitrary shape and abstract away from the actual geometry of the space. Topological information (cohomology generators and their intersections with paths) sufficiently represents the space. Additionally, the usage of algebraic topology makes our method dimension-independent, which extends the area of applications.

Acknowledgments. P.D. was partially supported by MNiSW grant N N206 625439, H.W. by the Austrian Science Fund (FWF) grant no. P20134-N13 and Foundation for Polish Science IPP Programme "Geometry and Topology in Physical Models". P.D. and H.W. are partially funded by the Google Research Awards Program.

References

1. Tovar, B., Cohen, F., LaValle, S.M.: Sensor Beams, Obstacles, and Possible Paths. In: Proc. Workshop on the Algorithmic Foundations of Robotics, WAFR (2008)
2. de Silva, V., Ghrist, R.: Coverage in sensor networks via persistent homology. Alg. & Geom. Topology 7, 339–358 (2007)
3. Soro, S., Heinzelman, W.: A Survey of Visual Sensor Networks. Advances in Multimedia 2009 (2009)
4. Edelsbrunner, H.: Geometry and Topology for Mesh Generation. Cambridge University Press, Cambridge (2001) ISBN 9780521793094
5. Hatcher, A.: Algebraic Topology. Cambridge University Press, Cambridge (2002)
6. Dłotko, P., Specogna, R.: Efficient cohomology computation for electromagnetic modeling. CMES: Computer Modeling in Engineering & Sciences 60(3), 247–278 (2010)
7. Dłotko, P., Specogna, R.: Critical analysis of the spanning tree techniques. SIAM Journal of Numerical Analysis (SINUM) 48(4), 1601–1624 (2010)
8. Cromwell, P.R., Beltrami, E., Rampichini, M.: The Borromean Rings. Mathematical Intelligencer 20(1), 53–62 (1998)

MAESTRO: Making Art-Enabled Sketches through Randomized Operations

Subhro Roy[1], Rahul Chatterjee[1], Partha Bhowmick[1], and Reinhard Klette[2]

[1] Indian Institute of Technology, Kharagpur, India
[2] The University of Auckland, New Zealand
{subhroroy.iitkgp,rahuliitkgp08,bhowmick}@gmail.com,
r.klette@auckland.ac.nz

Abstract. Contemporary digital art has an overwhelming trend of non-photorealism emulated by different algorithmic techniques. This paper proposes such a technique that uses a randomized algorithm to create artistic sketches from line drawings and edge maps. A *curve-constrained domain* (CCD) is defined by the *Minkowski sum* of the input drawing with the structuring element whose size varies with the pencil diameter. Each curve segment is randomly drawn in the CCD in such a way that it never intersects itself, whilst preserving the overall input shape. An artist's usual trait of making irregular strokes and sub-strokes with varying shades while sketching, is realistically captured in this randomized approach. Simulation results demonstrate its efficacy and elegance.

1 Introduction

Non-photorealistic rendering, originated as a promising digital art about two decades back [CAS97, VG91, VB99], has gained significant impetus in recent times [Deu10, GG01, LMHB00, MG02, Mou03, RMN03]. The works are mostly based on simulating the physical model through frictional coefficient, viscosity, smear factors, force factors, etc. [KHCC05, KNC08, KCC06, OSSJ09, PSNW07]. The factors of irregularity and obscurity arising out of an artist's creative mind—which prevail in an artistic creation and hence differentiate it from a machine-generated product—are, however, seldom noticed in the existing approaches. In fact, unless some (artistic) randomization is imparted, it is practically impossible to simulate an artistic creation, since the mystical, fanciful mind of an artist can hardly be scientifically modelled.

To incorporate a randomization factor while sketching a figure out of a set S of (irreducible) digital curve segments, corresponding to a real-world object, a novel simulation technique is proposed (Fig. 1). The fact that an artist often uses irregular strokes and sub-strokes with varying shades while sketching is rightly captured in our randomized curve sketching. Some sketched segments get lightly shaded in our technique compared to other heavily-shaded ones, and a sub-segment also may be lighter or deeper in shade compared to the rest of the segment, which characterizes the novelty of our algorithm.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 318–326, 2011.

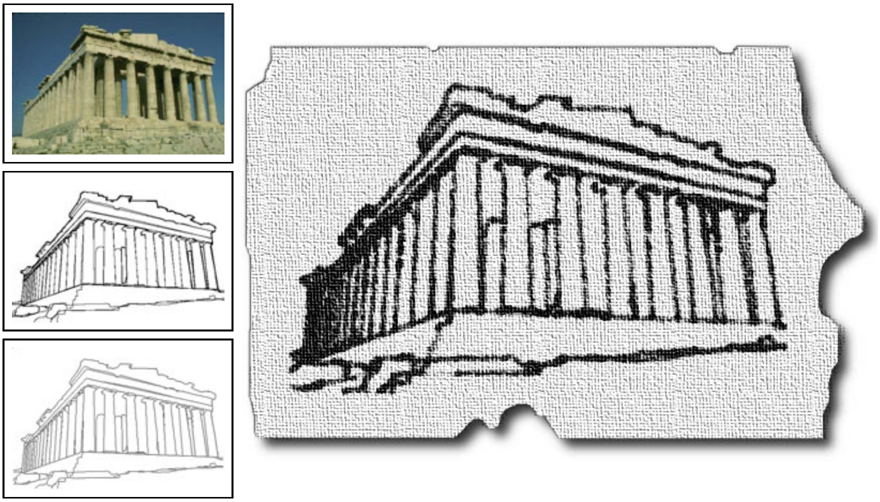


Fig. 1. Proposed algorithm. *Left-top:* Input image. *Left-mid:* After edge detection. *Left-bottom:* Skeletal image. *Right:* Final output by our algorithm, which resembles a crayon-drawn line sketch on a piece of handmade paper.

Preliminaries. A linear-time algorithm to generate random digital curves in a closed canvas is proposed recently in [BPR10]. The work proposed here rests on the same theoretical foundation, but extends the algorithm further for drawing random curves in CCD. Both the grid-point model and the cell model in $\mathbb{Z}^2$ are adopted in our work for their theoretical correspondence [KR04].

The input (thinned) digital image is first decomposed into a set $S = \{C_h\}_{h=1}^{N}$ of digital curves, each of which is simple and irreducible [KR04]. For each curve C_h, we prepare its *curve-constrained domain* (CCD), namely $\mathbb{D}_h$, in which the segment (corresponding to C_h) sketched by the pencil will lie. Figure 2 illustrates a simple case where the segment starts from the point p and ends at the point q. Notice that the points p and q lie in two cells of $\mathbb{D}_h$. Vertices and centres of cells in $\mathbb{D}_h$ are assumed to be grid points in $\mathbb{Z}^2$, also simply called *points* for brevity in this paper. For each point $p \in C_h$, we take its Minkowski sum [KR04], namely $\mathbb{M}_p = \{q : q \in \mathbb{Z}^2 \wedge ||p - q|| \leq \lfloor t/2 \rfloor\}$, t being the width/thickness of the pencil-tip; then $\mathbb{D}_h$ is defined by $\bigcup\limits_{p \in C_h} \mathbb{M}_p$.

2 Curve Randomization in CCD

Contrary to polygon-generation algorithms [ZSSM96, AH96] that work with input vertices generated randomly but *a priori*, our algorithm generates new points *on the fly* (also called *online* in [KR04]) while creating a digital curve ρ. The curve ρ starts from $p = p_1$ and randomly chooses all the successive points, eventually ending at the destination point q (Fig. 2). The difficulty lies in making ρ *one pixel wide everywhere without intersecting itself*, thus becoming irreducible and

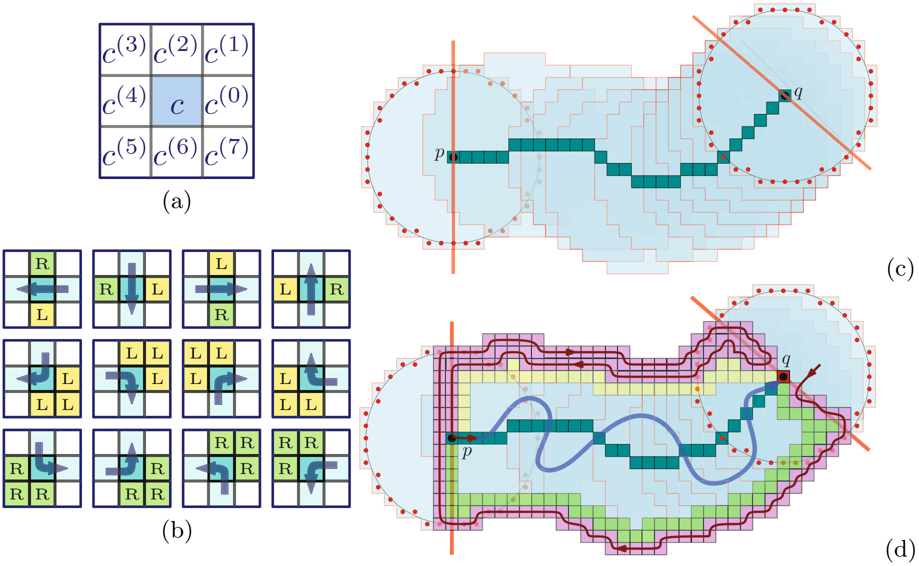


Fig. 2. (a) Neighborhood of a cell c: $A_0(c) = \{c^{(0)}, c^{(1)}, \ldots, c^{(7)}\}$; $A_1(c) = \{c^{(0)}, c^{(2)}, c^{(4)}, c^{(6)}\}$; $N_\alpha(c) = A_\alpha(c) \cup \{c\}$, $\alpha \in \{0, 1\}$. (b) Three types of turns with four combinatorial cases each. The current cell c_i is shown in blue, and the previous and the next cells in faded blue. (c) Minkowski sum (in blue) of a typical (simple and irreducible) digital curve (deep green) from p to q; red lines show the borderlines through p and q for CCD initialization. (d) Cells occupied ($\beta > 0$) by the initialized curve are shown in violet.

simple. This calls for detecting every possible "narrow-mouthed" *trap* formed by the previously generated part of ρ, which, if entered into, cannot be exited without touching or intersecting ρ.

Parameters and Principle. A cell c (of a CCD, say, $\mathbb{D}_h$) is said to be *occupied* if and only if the generated part of curve ρ already passes through c; otherwise it is *free*. We use the following parameters for a cell c (see Fig. 2):

The *blocking factor* $\beta(c)$ is a 5-bit number given by the combinatorial arrangement of the occupied and the free cells in $N_1(c)$. The most significant bit of $\beta(c)$ corresponds to c itself, and the other four bits correspond to the four cells lying right, top, left, and below of c in that order. If a cell in $N_1(c)$ is occupied, then the corresponding bit of $\beta(c)$ equals 1, otherwise 0. Thus, $\beta(c) = 0$ implies that ρ is not (yet) passing through any cell in $N_1(c)$. If $0 < \beta(c) < 16$, then c is free but one or more cells in $A_1(c)$ are occupied. If $\beta(c) \geq 16$, then c is occupied.

The *directional label* $\delta(c)$ is used if $0 < \beta(c) < 16$ which takes its value then from $\{\mathtt{L}, \mathtt{R}, \mathtt{B}\}$, with the interpretation: $\mathtt{L}$ = left, $\mathtt{R}$ = right, $\mathtt{B}$ = both left and right, depending on the position of c relative to the direction of traversal of ρ in the cell(s) of $A_0(c)$. We use $\mathtt{X}$ for the initialized value. While the construction

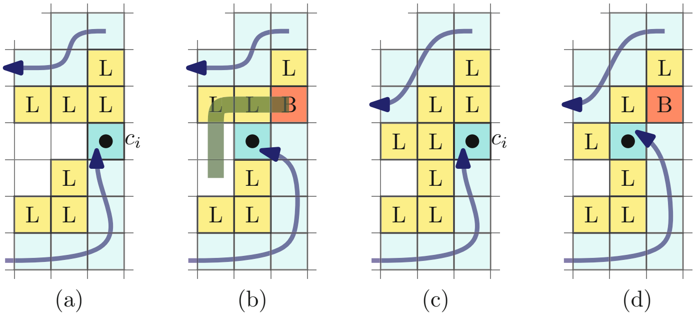


Fig. 3. Distinguishing the formation of a hole (a,b) from an ensuing hole (c,d). **Ensuing hole:** (a) Before formation, all the concerned cells have label `L`. (b) After formation, label of $b_i := c_i^{(2)}$ gets modified to `B`, and a free path exists from each free cell in $E_i \cap N_4(c_{i+1}) := \{c_{i+1}^{(2)}, c_{i+1}^{(4)}, c_{i+1}^{(6)}\}$ to b_i. **Hole:** (c) Before formation, cells have label `L`. (d) After formation, label of $b_i := c_i^{(2)}$ becomes `B`, and a free path to b_i is not possible from $c_{i+1}^{(4)}$ and $c_{i+1}^{(6)}$, as $c_{i+1}^{(3)}$ is blocked.

of ρ is in progress, blocking factors and directional labels have interim values, which are updated and become final values when ρ is finished.

Initialization of CCD. The cells c_p and c_q, corresponding to p and q, are obtained first (Fig. 2). The initialized curve ρ enters $c_q^{(6)}$ and then progresses through the border cells, to finally reach the cell c_p. By this initialization, c_q is free and has `B` as δ-value, whereas all other border cells are occupied, the (actual) random curve starts from p, and the free cells, adjacent to the border cells, have `L` or `R` as δ-value. While generating the random curve, if some cell c is visited which is adjacent to some border cell, then the corresponding parameters of c are updated accordingly. The initialized and the runtime parameters help advancing the curve in a random-yet-'safe' direction. Clearly, that *virtual* part of ρ lying in the border cells of $\mathbb{D}_h$ is not random, and hence not considered as being a part of the random curve.

Progressing the Random Curve. The *current cell*, which ρ has currently entered, is denoted by c_i $(i > 1)$, unless mentioned otherwise. The cell c_i corresponds to the ith iteration of our algorithm. Parameters β and δ are updated in (appropriate cells of) $A_0(c_i)$, as shown in Fig. 2. Each current cell c_i has a *previous cell*, c_{i-1}, from where ρ has entered c_i, and a *next cell*, c_{i+1}, where ρ will enter next. The cells belonging to the region $\widetilde{N}(c_i) := A_0(c_i) \smallsetminus (A_1(c_{i-1}) \cup A_1(c_{i+1}))$ are labelled in the ith iteration, as illustrated in Fig. 2.

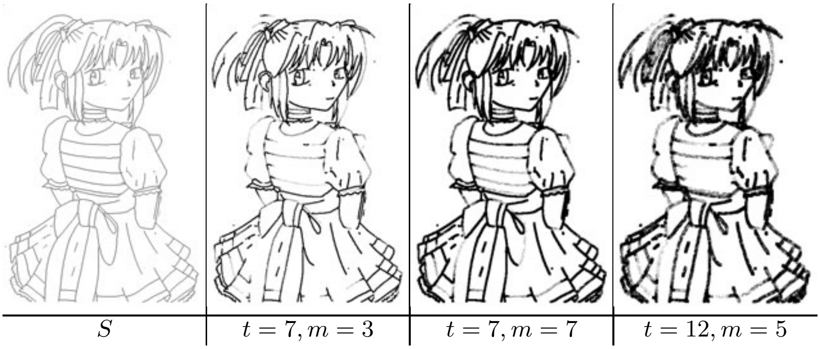
From the current cell c_i, the next cell c_{i+1} is (randomly) chosen in such a way that there exists at least one *free path* from c_{i+1} to the destination cell c_q. (A free path from a cell c_i to a cell c_{i+k}, $k > 1$, is given by a sequence of cells, $\rho(c_i, c_{i+k}) := \langle c_i, c_{i+1}, \ldots, c_{i+k} \rangle$, such that each cell in $\langle c_{i+1}, \ldots, c_{i+k-1} \rangle$ is free and distinct, and every two consecutive cells in $\rho(c_i, c_{i+k})$ are 1-adjacent.) A *safe edge* of c_i is a possible exit edge; the algorithm selects randomly one of the safe edges for exit. For the current cell c_i we have the *free region* R_i of all free cells c of $\mathbb{D}_h$ such that there exists still at least one free path from c to c_q. Similarly, a *blocked region* H is a maximal (connected) region of free cells such that there does not exist any free path from any cell of H to c_q. A cell in H is said to be *blocked*, and edges of a blocked cell are also *blocked*. There exists a free path from the current cell c_i to the destination cell c_q if and only if $A_1(c_i) \cap R_i \neq \emptyset$. (If $A_1(c_i) \cap R_i \neq \emptyset$, then there exists a free cell $c_i^{(t)} \in A_1(c_i)$ lying in R_i. Conversely, the existence of a free path from c_i to c_q implies that at least one cell of $A_1(c_i)$ is in R_i, thus $A_1(c_i) \cap R_i \neq \emptyset$.) As a result, the edge between c_i and $c_i^{(t)}$ is safe if and only if $c_i^{(t)}$ belongs to R_i.

Ensuring Simple and Irreducible Property: ρ is allowed to enter and exit a cell at most once. Hence, an exit edge of the current cell c_i cannot be an entry edge of the next cell if the latter is already occupied (using $\beta(c_i)$). Furthermore, a blocked edge cannot be an exit edge. The crux of the problem is, therefore, to decide whether or not an edge of c_i is a blocked edge. Each event of forming a hole is detected based on (changes in the components of the cells in) $A_0(c_i)$. The advantage of detecting such a *hole event* is that, once ρ enters the next cell c_{i+1} from c_i by selecting a safe edge, it can never enter the hole H formed by c_i, since H gets surrounded by occupied cells after it is formed. Further characterization of cells in the local neighbourhood of c_i are required to distinguish whether there is a hole event or an event of an *ensuing hole* (Fig. 3). $E_i \subset R_i$ defines an ensuing hole corresponding to c_i if and only if

(e1) there exists $c \in \widetilde{N}(c_i)$ such that $\delta(c, i) = \mathtt{B}$,
(e2) for each $c' \in E_i$, we have that $\delta(c', i) \in \{\mathtt{L}, \mathtt{R}, \mathtt{X}\}$,
(e3) there exists a free path $\rho(c_{i+1}, c_q)$, and for any such path, c is on $\rho(c_{i+1}, c_q)$.
Note that, $\delta(c, i)$ denotes the label of cell c when the current cell is c_i.

Either a hole or an ensuing hole is created if and only if at least one free cell in $\widetilde{N}(c_i)$ gets the label $\mathtt{B}$ as c_i becomes the current cell. The current cell c_i gives rise to an ensuing hole E_i if and only if there exists a free cell $b_i \in \widetilde{N}(c_i)$ with
(E1) $\delta(b_i, i) = \mathtt{B}$;
(E2) there exists $\rho(a_i, b_i) \subseteq A_0(c_{i+1})$ for each $a_i \in E_i \cap A_1(c_{i+1})$. In particular, c_i gives rise to a hole H_i if and only if E1 is true and E2 is false. The proof follows from the combinatorial arguments given in [BPR10].

Final Sketch Creation. For each curve C_h in S, we create m random curves. Note that, S is obtained in our work by Canny edge detection [Can86] and thinning [RK82]. If p and q be the respective start and end points of the curve C_h, then each of these m random curves is made to start from p and end at q. Further, due to the curve-constrained domain, $\mathbb{D}_h$, corresponding to C_h, each random curve strictly lies in $\mathbb{D}_h$. The cells of $\mathbb{D}_h$ are always (re-)initialized for creating each instance of the m random curves corresponding to C_h.



S | $t = 7, m = 3$ | $t = 7, m = 7$ | $t = 12, m = 5$

Fig. 4. Effect of varying m versus t ($\gamma_{\max} = 255, \gamma_0 = 0$)

Let $S_h = \left\{ C_h^{(z)} \right\}_{z=1}^{m}$ be the set of m random curves corresponding to C_h. Let c_h be a cell of the domain $\mathbb{D}_h$. We maintain a counter, namely $\textsc{count}[c_h]$, corresponding to each $c_h \in \mathbb{D}_h$. Each such $\textsc{count}[c_h]$ is initialized to 0 before generating the random curves in $\mathbb{D}_h$. Whenever a random curve $C_h^{(z)}$ visits c_h, $\textsc{count}[c_h]$ is incremented. Thus, after all m random curves are constructed in $\mathbb{D}_h$, we get $0 \leq \textsc{count}[c_h] \leq m \; \forall c_h \in \mathbb{D}_h$.

In order to create the artistic curve $\tilde{C}_h$ corresponding to C_h, we use the counter values $\{\textsc{count}[c_h] : c_h \in \mathbb{D}_h\}$. For each $c_h \in \mathbb{D}_h$, the corresponding image pixel is intensified to the value

$$\gamma_{\max} - \left(\gamma_0 + \frac{\textsc{count}[c_h]}{m} \times (\gamma_{\max} - \gamma_0) \right),$$

since we consider 8-bit intensity of the image (with $\gamma_{\max} = 255, \gamma_0 = 0$) as the final output corresponding to the input set S. To achieve an overall darker intensity (as in Fig. 4) in simultaneity with the randomized finish, we scale the colour spectrum to a smaller interval, namely $[\gamma_0, \gamma_{\max} - \gamma_0]$, and assign the pixel intensity by setting γ_0 to an appropriately high value.

3 Results and Conclusion

We have developed the software in Java™ API, version 1.5.2, the OS being Linux Fedora Release 7, Kernel version 2.6.21.1.3194.fc7, Dual Intel Xeon Processor 2.8 GHz, 800 MHz FSB, and have tested it on various digital images. Snapshots on a typical set are already given in Fig. 1. A summary of results for a few images presented in Table 1 shows that as the width of pencil-tip increases, the runtime also increases, since it needs a larger number of iterations to create sufficient stroke intensity.

Fig. 4 shows the effect of number of iterations m with changing width of the pencil-tip, t. Clearly, for a given value of t, the stroke intensity increases with increase in m. An appropriate combination of m and t is, therefore, required to

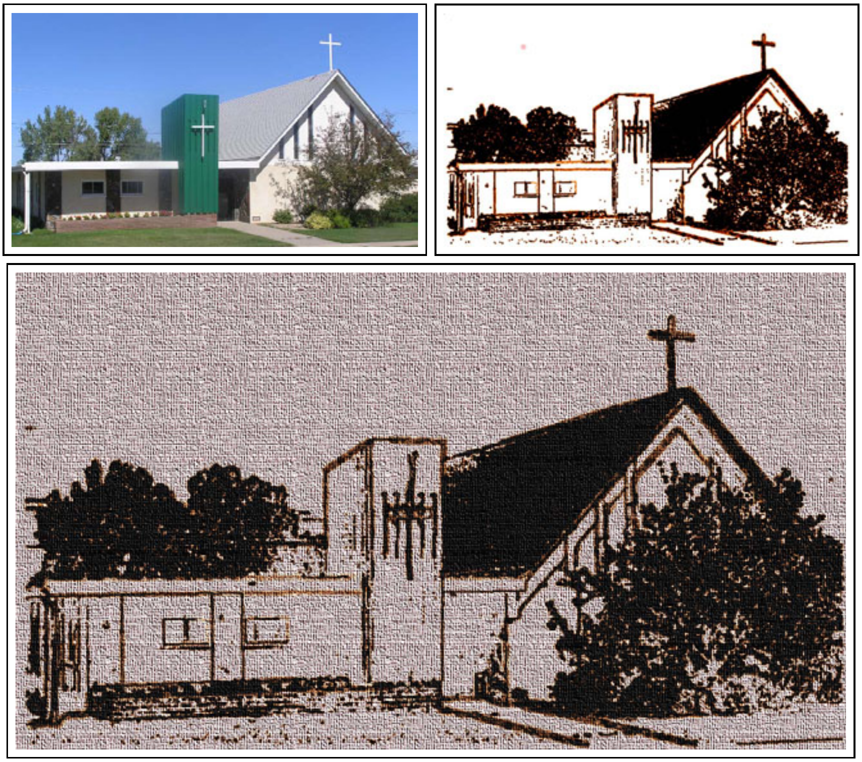


Fig. 5. Results on another image. Top-left: A photograph. Top-right: Product of our algorithm. Bottom: After overlaying on a canvas.

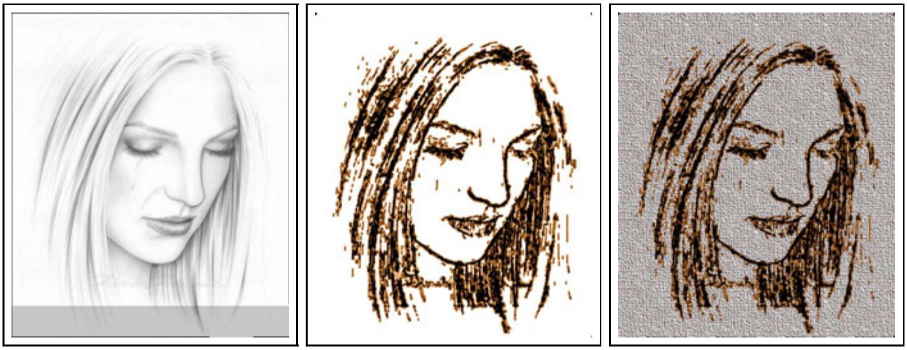


Fig. 6. Effect of using mixed pencils. Left: Input image. Middle: Product of our algorithm. Right: After overlaying on a canvas. Note that our algorithm uses thick curves in the relevant portion (e.g., nose) and thin curves for small details, which creates the desired artistic touch.

Table 1. Summary of simulation results

Image	w	h	n	t	m	T
houses	480	320	3441	8	10	10.871
houses	480	320	3441	5	5	6.257
nestle	320	480	2579	8	10	9.755
elephants	480	320	6256	5	5	10.306
vase	220	400	5048	5	5	3.333

w = image width; h = image height; n = number of curve points in the input image; t = width/thickness of pencil-tip; m = number of random curves; T = CPU time in seconds for the algorithm to produce the final output.

achieve the aesthetic quality. Finer details can be captured with a fine-tipped pencil (low value of t), as presented in Fig. 4. Figures 5 and 6 show how our algorithm successfully produce the desired artistic impression—whether the type of input be a line-sketch or a photograph. Figure 6 also shows the usage of mixed pencils to take care of different regions of interest. For a fairly long curve that possibly signifies a strong structural information, demanding a bold stroke from the artist, a thick and bold line is sketched by our algorithm. The non-uniformity of shade gives a crayon-like appeal, thus creating an artistic finish.

References

[AH96] Auer T., Held M.: Heuristics for the generation of random polygons. In: Proc. CCCG, pp. 38–44 (1996)

[BPR10] Bhowmick P., Pal O., Klette R.: A linear-time algorithm for generation of random digital curves. In: Proc. PSIVT 2010, pp. 168–173 (2010)

[Can86] Canny, J.: A computational approach to edge detection. IEEE Trans. PAMI 8(6), 679–698 (1986)

[CAS97] Curtis, C.J., Anderson, S.E., Seims, J.E., Fleischer, K.W., Salesin, D.H.: Computer-generated watercolor. In: Proc. SIGGRAPH 1997, pp. 421–430 (1997)

[Deu10] Deussen, O.: Oliver's artistic attempts (random line) (2010), http://graphics.uni-konstanz.de/artlike

[GG01] Gooch, B., Gooch, A.: Non-photorealistic rendering. A.K. Peters Ltd., NY (2001)

[KCC06] Kang, H.W., Chui, C.K., Chakraborty, U.K.: A unified scheme for adaptive stroke-based rendering. The Vis. Computer 22, 814–824 (2006)

[KHCC05] Kang, H.W., He, W., Chui, C.K., Chakraborty, U.K.: Interactive sketch generation. The Visual Computer 21, 821–830 (2005)

[KNC08] Kopf, J., Neubert, B., Chen, B., Cohen, M., Cohen-Or, D., Deussen, O., Uyttendaele, M., Lischinski, D.: Deep photo: Model-based photograph enhancement and viewing. In: SIGGRAPH Asia 2008, pp. 1–10 (2008)

[KR04] Klette, R., Rosenfeld, A.: Digital Geometry: Geometric Methods for Digital Picture Analysis. Morgan Kaufmann, San Francisco (2004)

[LMHB00] Lake, A., Marshall, C., Harris, M., Blackstein, M.: Stylized rendering techniques for scalable real-time 3d animation. In: Proc. NPAR 2000, pp. 13–20 (2000)

[MG02] Majumder, A., Gopi, M.: Hardware accelerated real time charcoal rendering. In: Proc. NPAR 2002, pp. 59–66 (2002)

[Mou03] Mould, D.: A stained glass image filter. In: Proc. EGRW 2003, pp. 20–25 (2003)

[OSSJ09] Olsen, L., Samavati, F.F., Sousa, M.C., Jorge, J.A.: Sketch-based modeling: A survey. Computers and Graphics 33(1), 85–103 (2009)

[PSNW07] Pusch, R., Samavati, F., Nasri, A., Wyvill, B.: Improving the sketch-based interface: Forming curves from many small strokes. The Visual Computer 23(9), 955–962 (2007)

[RK82] Rosenfeld, A., Kak, A.C.: Digital Picture Processing, 2nd edn. Academic Press, NY (1982)

[RMN03] Rudolf, D., Mould, D., Neufeld, E.: Simulating wax crayons. In: PG 2003, pp. 163–172 (2003)

[VB99] Verevka, O., Buchanan, J.W.: Halftoning with image-based dither screens. In: Proc. Graphics Interface 1999, pp. 167–174 (1999)

[VG91] Velho, L., Gomes, J.d.M.: Digital halftoning with space filling curves. In: Proc. SIGGRAPH 1991, pp. 81–90 (1991)

[ZSSM96] Zhu C., Sundaram G., Snoeyink J., Mitchell J. S. B.: Generating random polygons with given vertices. Computational Geometry Theory and Applications, 277–290 (1996)

Improved Working Set Selection for LaRank

Matthias Tuma[1,⋆] and Christian Igel[2]

[1] Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany
matthias.tuma@rub.de

[2] Department of Computer Science, University of Copenhagen, Denmark
igel@diku.dk

Abstract. LaRank is a multi-class support vector machine training algorithm for approximate online and batch learning based on sequential minimal optimization. For batch learning, LaRank performs one or more learning epochs over the training set. One epoch sequentially tests all currently excluded training examples for inclusion in the dual optimization problem, with intermittent *reprocess* optimization steps on examples currently included. Working set selection for one reprocess step chooses the *most violating pair* among variables corresponding to a random example. We propose a new working set selection scheme which exploits the gradient update necessarily following an optimization step. This makes it computationally more efficient. Among a set of candidate examples we pick the one yielding *maximum gain* between either of the classes being updated and a randomly chosen third class. Experiments demonstrate faster convergence on three of four benchmark datasets and no significant difference on the fourth.

1 Introduction

Support vector machines (SVMs, e.g. [1]) have attractive theoretical properties and give good classification results in practice. Training times between quadratic and cubic in the number of training examples however impair their applicability to large-scale problems. Over the past years well-performing *online* variants of binary and multi-class SVM solvers have been proposed and refined [2,3,4,5,6]. Online SVMs can be preferable to standard SVMs even for batch learning. On large datasets that prohibit calculating a close to optimal solution to the SVM optimization problem, online SVMs can excel in finding good approximate solutions quickly. The prominent online multi-class SVM LaRank was introduced by Bordes, Bottou, Gallinari, and Weston [4]. It relies on the multi-class SVM formulation proposed by Crammer and Singer (CS, [7]). We refer to LaRank-like solvers [2,4] as *epoch-based* since they complete one or more epochs over a training set, aiming at well-performing hypotheses after as few as a single epoch.

When using universal, non-linear kernels – which are a prerequisite for consistency, cf. [8] – sequential minimal optimization (SMO, [9]) solvers are the method of choice for SVM training. Their time requirements strongly depend

[⋆] M.T. acknowledges a scholarship by the German National Academic Foundation.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 327–334, 2011.

on stopping criteria, working set selection, and implementation details such as kernel cache organization, shrinking, etc. This paper focuses on an improvement to working set selection for SMO steps in LaRank.

Section 2 formally introduces the CS machine. We give a general definition of epoch-based CS solvers and restate the LaRank algorithm [4]. One of LaRank's building blocks is the random selection of examples for *reprocess*-type optimization steps. Random example selection provides the advantage of a constant time operation, however at the risk of conducting an optimization step yielding little gain for the overall problem. We propose an alternative example selection scheme which is both gain-sensitive and can be carried out fast enough to speed up the overall convergence of dual, primal, and test error. Empirical evaluations are presented in Section 3, followed by our conclusions and outlook in Section 4.

2 Multi-class SVMs

Consider an input set X, an output set $Y = \{1, \dots, d\}$, a labeled training set $S_N = \{(x_i, y_i)\}_{1 \leq i \leq N} \in (X \times Y)^N$ of cardinality N, and a Mercer kernel [1] function $k : X \times X \to \mathbb{R}$. Then a trained all-in-one multi-class SVM without bias assigns to an example $x \in X$ the output class label

$$h(x) = \arg\max_{y \in Y} \sum_{i=1}^{N} \beta_i^y k(x, x_i) \ . \tag{1}$$

Training the SVM is equivalent to determining the parameter vector $\beta \in \mathbb{R}^{dN}$. Its component β_i^y constitutes the contribution of example i to the kernel expansion associated with class y. Iff $\exists y \ \ \beta_i^y \neq 0$, we say that i or β_i is a support pattern, and iff $\beta_i^y \neq 0$, we say that y or β_i^y is a support class for sample i.

2.1 Crammer-Singer Type Multi-class SVMs

Following the notation in [4] and given a regularization parameter $C \in \mathbb{R}^+$, Crammer-Singer type multi-class SVMs [7] determine the parameter vector β by solving the dual optimization problem

$$\max_{\beta} \quad \sum_i \beta_i^{y_i} - \frac{1}{2} \sum_{i,j} \sum_y \beta_i^y \beta_j^y k(x_i, x_j) \tag{2}$$

$$\text{s.t.} \quad \forall i \ \ \forall y \ \ \beta_i^y \leq C\delta(y, y_i) \tag{3}$$

$$\forall i \ \ \sum_y \beta_i^y = 0 \ , \tag{4}$$

with Kronecker delta δ. The derivative of (2) w.r.t. the variable β_i^y is given by

$$g_i^y = \delta(y, y_i) - \sum_j \beta_j^y k(x_i, x_j) \ . \tag{5}$$

Two notable consequences arise from the specific form of problem (2). Constraint (4) in practice restricts SMO solvers to working sets of size two with both working variables corresponding to the same training example. This closely links working set selection to example selection. Second, because the quadratic part of problem (2) is entirely composed of diagonal sub-matrices, altering a variable β_i^c only propagates through to gradients g_j^c involving the same class label.

2.2 Epoch-Based Crammer-Singer Solvers

We understand an epoch-based Crammer-Singer (EBCS) solver to carry out optimization epochs over a training set according to Alg. 1. Specific variants of EBCS solvers are then realized through different implementations of the sub-algorithms WSS_{new}, WSS_{rep}, and R. These control working set selection for (i) non-support patterns and (ii) support patterns, as well as (iii) the relative ratio between optimization steps on non-support and support patterns, respectively. All three will in practice depend on the joint state of both solver and solution. Note that we understand the SMO steps in lines 7 and 11 of Alg. 1 to potentially leave β unaltered, for example if both variables are actively constrained.

Algorithm 1. Epoch-based Crammer-Singer solver

```
   Input: training set $S_N$, epoch limit $e_{max}$, working set selection
          algorithms $\text{WSS}_{\text{new}}$ and $\text{WSS}_{\text{rep}}$, step selection algorithm R
 1 $\beta \leftarrow 0$
 2 for $e \leftarrow 1$ to $e_{max}$ do                                  // 1 loop = 1 epoch
 3     Shuffle training set $S_N$ jointly with $\beta$
 4     for $i \leftarrow 1$ to $N$ do                                  // 1 loop = 1 sample
 5         if $\forall\, y \;\; \beta_i^y = 0$ then                            // process new sample
 6             Choose $(c, e) \in Y^2$ according to $\text{WSS}_{\text{new}}$
 7             SMO-step on $(\beta_i^c, \beta_i^e)$ and gradient update
 8             while not R do                              // reprocess old samples
 9                 Choose $(j, c, e) \in \{1, \ldots, N\} \times Y^2$ according to $\text{WSS}_{\text{rep}}$
10                 if $\exists y \;\; \beta_j^y \neq 0$ then
11                     SMO-step on $(\beta_j^c, \beta_j^e)$ and gradient update
```

LaRank. The popular EBCS solver LaRank [4] in its query to WSS_{rep} (Line 9 of Alg. 1) chooses the example index j randomly. For both WSS_{rep} and WSS_{new}, the class indices (c, e) are selected according to the *most violating pair* heuristic (MVP, [10]) on the given example. In addition, WSS_{rep} operates in two different modes, WSS_{old} and WSS_{opt}, which perform MVP among all classes or all support classes of one example, respectively. The resulting three step variants *processNew*, *processOld*, and *processOpt* are chosen from in a stochastic manner. Their probabilistic weights are adapted through three slowly relaxing linear dynamical systems with attractors to the current dual gain rate of each step variant. For alternative, deterministic step selection schemes also see [2,5,6].

Algorithm 2. Gain-sensitive working set selection for LaRank

```
SMO-step on $(\beta_i^c, \beta_i^e)$
$(w_{max}, v) \leftarrow (0, \emptyset)$
Pick random class $t$, $t \notin \{c, e\}$
for $(j, y) : \beta_j^y \neq 0,\ y \in \{c, e\}$ do          // loop through support classes
    Update $(g_j^y)$
    if $\beta_j^t \neq 0$ then
        $w \leftarrow$ clipped SMO-Gain$(\beta_j^y, \beta_j^t)$
        if $w > w_{max}$ then                              // found new best candidate
            $(w_{max}, v) \leftarrow (w, j)$
if $w_{max} = 0$ then                                      // fallback to random
    $v \leftarrow$ index of random support pattern
Provide example $v$ upon next call to WSS$_{\text{rep}}$
```

Gain-sensitive working set selection. LaRank and its binary predecessor LaSVM [2] are inspired by perceptron-like kernel machines. As such, the random traversal of hitherto excluded training samples around line 4 of Alg. 1 is conceptually well-founded. Since first and second order working set selection coincide for CS, MVP can further be seen as a viable approximation to clipped gain working set selection [11,3]. Another relevant building block of EBCS solvers is the example selection procedure for WSS$_{\text{rep}}$. A naive deterministic alternative to LaRank's random selection scheme would be to compute the full $\arg\max_i \arg\max_{(c,e)}$ of the clipped or unclipped gain. Yet, the computational effort outburdens the potential gain, especially if, as for original LaRank, not all gradients are being cached. The LaRank algorithm with minimal cost and random gain can thus be seen as lying on one end of all possible example selection methods and the argmax-scheme with maximum cost and maximum gain on the other. This paper explores the question whether the already well-performing LaRank algorithm can be further improved by an example selection scheme for which the added cost (relative to instant example selection) is outweighed by the gain advantage received in turn (relative to the average gain of MVP on random examples).

We propose to exploit the gradient update necessarily following each SMO step to select the next "old" example. Similar to [11], reusing information recently computed promises efficient working set selection. Let (β_i^c, β_i^e) be the pair of variables altered by the last SMO step. Then, according to (5), the subset of all gradients $\{g_j^c, g_j^e\}_{1 \leq j \leq N}$ currently stored by the solver must be updated. For LaRank these are all g_j^y, $y \in \{c, e\}$, for which $\beta_j^y \neq 0$. As the solver looks at this subset in any case, it suggests itself to select the next old example according to some property of all gradients being updated. We propose as such a property the clipped gain achievable by a SMO step between the variable β_j^y the gradient g_j^y of which is being updated and, fixed within each update loop, a random third class t. If β_j^t is not a support class, it is not considered. Alg. 2 summarizes the resulting example selection procedure following both SMO steps in lines 7 and 11 of Alg. 1. If no feasible pair can be identified, a fallback to a random sample is guaranteed. In practice, this only occurs in the first few iterations. After Alg. 2,

Table 1. Datasets, SVM hyperparameters, and average reprocess step rates

	Train Ex.	Test Ex.	Classes	Features	C	$k(x,z)$	s_{old}	s_{opt}
USPS	7291	2007	10	256	10	$e^{-0.025(x-z)^2}$	1.94	36.7
LETTER	16000	4000	26	16	10	$e^{-0.025(x-z)^2}$	1.75	82.6
INEX	6053	6054	18	167295	100	$x \cdot z$	3.67	35.1
MNIST	60000	10000	10	780	1000	$e^{-0.02(x-z)^2}$	1.65	53.4

a call to WSS_{opt} will directly return (β_v^y, β_v^t), while a call to WSS_{old} returns the MVP within the candidate example v. In the rare case that the latter does not yield a feasible variable pair, we also choose a random example in the next step.

Compared to the original version, Alg. 2 adds the computational burden of checking whether $\beta_j^t = 0$ for all examples for which $\beta_j^y \neq 0$, $y \in \{c, e\}$. For those examples for which $\beta_j^t \neq 0$ we say that we have a *hit* between class y and t. For every hit the potential gain of a SMO step on (β_j^y, β_j^t) has to be calculated and compared to the current maximum candidate. Because for each class the support classes lie sparse in the total set of support patterns, the gain calculation is only conducted in a fraction of update steps. Yet still, experiments not documented here indicate that Alg. 2 does not typically improve upon LaRank. Since Alg. 2 is carried out after each SMO step, the resulting constant time cost propagates through to all three average gain rates which steer the stochastic step selection procedure. The added time is negligible for the more costly step types *processNew* and *processOld*, but large enough to make the selection of *processOpt* significantly more unlikely. This in turn impedes the removal of useless support patterns, which again makes update steps more costly.

We reduce the computational cost by entering candidate examination at line 10 of Alg. 2 only for a subset of all variables being updated. In detail, we introduce a parameter D representing the desired number of hits within the entire update loop. Starting from a random index we only enter candidate examination at line 10 while less than D hits have occurred. Note that the best of D hits with probability $1 - x^D$ is better or equal to the best in a fraction x of all possible hits (e.g., theorem 6.33 in [1]). We choose $D = 10$, for which the probability of the best of D random hits being in the highest quintile of all possible hits is $\sim 90\%$, and divide these ten hits evenly between the two classes being updated. We further provide an incentive towards sparser solutions and hence shorter gradient update times by slightly modifying line 12 of Alg. 2. If a SMO step on a candidate hit would eliminate at least one of the two support classes, that step is given preference over a non-eliminating candidate step. Between candidates of identical priority the resulting gain remains the selection criterion, just as stated in line 12 of Alg. 2. For brevity we refer to this final algorithm employing gain sensitive example selection in LaRank reprocess steps as “GaLa”.

3 Experiments

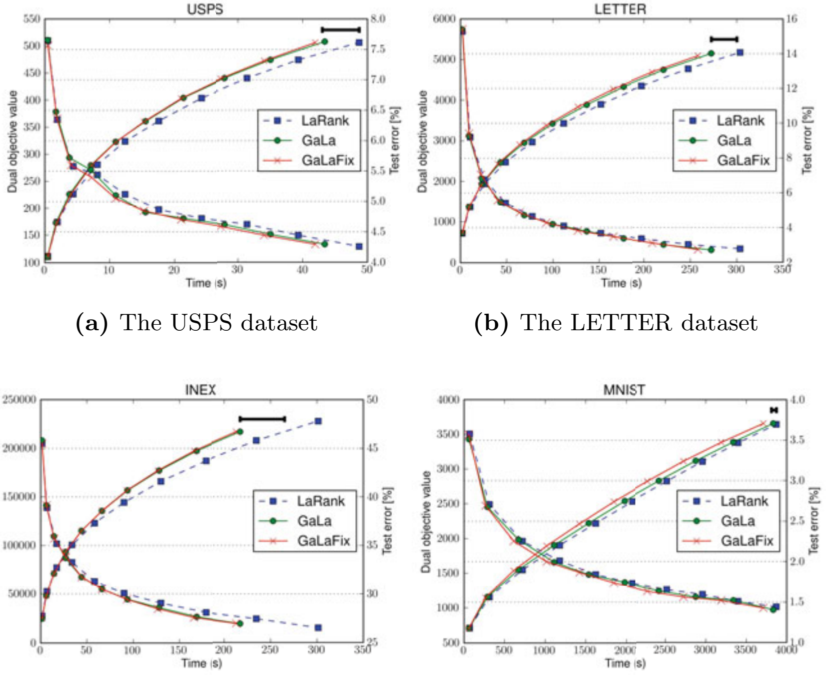
We incorporated GaLa into the original LaRank implementation and conducted our comparison on the original benchmark datasets, both obtainable at the software page of [4]. Tb. 1 lists the corresponding dataset characteristics and SVM hyperparameters.[1] We further wish to rule out that our results are merely an artifact of GaLa nudging the stochastic step selection mechanism to for some reason more suitable relative step rates. We therefore besides LaRank and GaLa considered a third variant for comparison, GaLaFix, in which we fixed the average number of *processOld* and *processOpt* steps per *processNew* in GaLa to those exhibited by LaRank. In detail, we for each dataset simultaneously let one LaRank and two dummy GaLa runs perform ten independent single-epoch trials and noted the average relative step rates $(\mathrm{s}_{\mathrm{old}}, \mathrm{s}_{\mathrm{opt}})$ of LaRank in Tb. 1. In the actual experiments we compare GaLaFix, clamped to these empirical step rates, to LaRank and GaLa, afterwards verifying that LaRank approximately reached the same step rates again. Fig. 1 shows the results obtained as mean averages over ten independent single-epoch trials on differently shuffled training sets. We for clarity excluded the primal training curves, which qualitatively follow those of the test errors. The horizontal black bar in the upper right of each plot illustrates the factual time advantage of GaLa over LaRank. It extends from the finish time of that method with lower final dual value to the linearly extrapolated time at which the respective other method reached the same dual value. Dividing the length of the line by the time of its later endpoint we note a speed-up of 12, 9, 18, and 2 percent for USPS, LETTER, INEX, and MNIST, respectively. Experiments were carried out on an eight-core 2.9 GHz machine with 8 MB CPU cache, 3.5 GB memory, using 500 MB of which as kernel cache, and no other avoidable tasks running besides all three methods in parallel.[2]

[1] As SVM hyperparameters were selected on the basis of "past experience" [4], the dual curve should probably be seen as most significant performance measure. Further, MNIST data and hyperparameters slightly vary between the printed and website version of [4], which we used and where the relevant differences are listed. We also slightly modified the LaRank implementation for training set shuffling, serialization, etc. The entire source code underlying our experiments can be obtained at http://image.diku.dk/igel/downloads.php. Besides the implementation described above, we added a complete re-implementation of an EBCS solver to the Shark machine learning library [12]. In that implementation one epoch of GaLa on MNIST on average reaches a dual of 3656 in 987 seconds, despite not speeding up sparse radial basis function kernel evaluations through precomputed norms as in the original.

[2] Similar to the note on the LaSVM software page [2] we observed performance variability across platforms also for LaRank. We ascribe this effect to the volatility of the step selection algorithm. E.g., if kernel computations are slightly *faster* on one machine, this will make *processNew* and *processOld* steps more likely, but might lead to an actual *decrease* in accuracy if the relative advantage for *processNew* is higher. Further, if the operation underlying gradient updates takes longer on one machine, this constant cost on all three step types will regularize the original step selection mechanism. In [5] the adaptive step selection mechanism is discarded for a deterministic one, at the cost of introducing an additional SVM hyperparameter.

3.1 Results and Discussion

For the first three datasets the proposed method arrives at the same dual values between 9% and 18% faster than the original approach. For the fourth dataset, MNIST, it only yields a marginal advantage of 2%. Possible reasons for this comparatively weak performance may be that the distribution of gradients is such that randomly picking an example holds no real disadvantage as compared to a gain-sensitive selection method. We also conducted minor experiments not documented here towards the role of the parameter D, but did not find qualitatively different results for reasonable changes in D. Third, it is notable that until around 2300 seconds, GaLaFix persistently sustains an advantage of 10 to 15% over LaRank. It might be enlightening to relate the subsequent decline to the onset of the kernel cache overflow, since that would most likely significantly perturb the target attractor for the probabilistic weight of *processOld* steps. This however is not straightforward as LaRank uses d class-wise kernel caches.



(a) The USPS dataset

(b) The LETTER dataset

(c) The INEX dataset

(d) The MNIST dataset

Fig. 1. Development of dual objective (left axis) and test error (right axis) of LaRank, GaLa, and GaLaFix over one epoch on four benchmark datasets

4 Conclusions

We proposed a gain-sensitive working set selection algorithm for LaRank by Bordes et al. [4], which is an epoch-based solver for the Crammer-Singer multi-class SVM [7]. Our new working set selection scheme improves learning speed and is conceptually compatible with a wide range of conceivable step selection procedures. While several approaches to step selection have been presented [2,4,5,6], a robust canonical solution has yet to be developed. We further believe that the method suggested here is a promising basis for parallelizing *processOpt* steps in LaRank. Since SMO steps and subsequent gradient updates are independent for disjunct class pairs, $d/3$ parallel SMO steps should with slight modifications be possible while still benefiting from gain-sensitive example selection.

References

1. Schölkopf, B., Smola, A.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2002)
2. Bordes, A., Ertekin, S., Weston, J., Bottou, L.: Fast kernel classifiers with online and active learning. Journal of Machine Learning Research 6, 1579–1619 (2005), http://leon.bottou.org/papers/bordes-ertekin-weston-bottou-2005
3. Glasmachers, T., Igel, C.: Second order SMO improves SVM online and active learning. Neural Computation 20(2), 374–382 (2008)
4. Bordes, A., Bottou, L., Gallinari, P., Weston, J.: Solving multiclass support vector machines with LaRank. In: Proceedings of the 24th International Conference on Machine Learning, pp. 89–96. OmniPress (2007), http://www-etud.iro.umontreal.ca/~bordesa/mywiki/doku.php?id=larank
5. Bordes, A., Usunier, N., Bottou, L.: Sequence labelling SVMs trained in one pass. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part I. LNCS (LNAI), vol. 5211, pp. 146–161. Springer, Heidelberg (2008)
6. Ertekin, S., Bottou, L., Giles, C.: Non-convex online support vector machines. IEEE Transactions on Pattern Recognition and Machine Intelligence 33(2), 368–381 (2011)
7. Crammer, K., Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research 2, 265–292 (2002)
8. Steinwart, I.: Support vector machines are universally consistent. Journal of Complexity 18, 768–791 (2002)
9. Platt, J.: Fast training of support vector machines using sequential minimal optimization. In: Schölkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods: Support Vector Learning, pp. 185–208. MIT Press, Cambridge (1999)
10. Joachims, T.: Making large-scale SVM learning practical. In: Schölkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods: Support Vector Learning, pp. 169–184. MIT Press, Cambridge (1999)
11. Glasmachers, T., Igel, C.: Maximum-gain working set selection for support vector machines. Journal of Machine Learning Research 7, 1437–1466 (2006)
12. Igel, C., Glasmachers, T., Heidrich-Meisner, V.: Shark. Journal of Machine Learning Research 9, 993–996 (2008), http://shark-project.sourceforge.net

Multi-task Learning via Non-sparse Multiple Kernel Learning

Wojciech Samek[1,2,*], Alexander Binder[1], and Motoaki Kawanabe[2,1,**]

[1] Technical University of Berlin, Franklinstr. 28 / 29, 10587 Berlin, Germany
{wojciech.samek,alexander.binder}@tu-berlin.de
[2] Fraunhofer Institute FIRST, Kekuléstr. 7, 12489 Berlin, Germany
motoaki.kawanabe@first.fraunhofer.de

Abstract. In object classification tasks from digital photographs, multiple categories are considered for annotation. Some of these visual concepts may have semantic relations and can appear simultaneously in images. Although taxonomical relations and co-occurrence structures between object categories have been studied, it is not easy to use such information to enhance performance of object classification. In this paper, we propose a novel multi-task learning procedure which extracts useful information from the classifiers for the other categories. Our approach is based on non-sparse multiple kernel learning (MKL) which has been successfully applied to adaptive feature selection for image classification. Experimental results on PASCAL VOC 2009 data show the potential of our method.

Keywords: Image Annotation, Multi-Task Learning, Multiple Kernel Learning.

1 Introduction

Recognizing objects in images is one of the most challenging problems in computer vision. Although much progress has been made during the last decades, performance of state-of-the art systems are far from the ability of humans. One possible reason is that humans do incorporate co-occurrences and semantic relations between object categories into their recognition process. On the contrary, standard procedures for image categorization learn one-vs-rest classifiers for each object class independently [2].

In this paper, we propose a two-step *multi-task learning (MTL)* procedure which can find out useful information from the classifiers for the other categories based on *multiple kernel learning (MKL)* [6], and its non-sparse extension [4]. In the first step we train and apply the classifiers independently for each class and construct extra kernels (similarities between images) from the outputs. In the second step we incorporate information from other categories by applying MKL with the extended set of kernels. Our approach has several advantages over standard MTL methods like Evgeniou *et al.* [3],

[*] né Wojcikiewicz

[**] We thank Klaus-Robert Müller for valuable suggestions. This work was supported by the Federal Ministry of Economics and Technology of Germany under the project THESEUS (FKZ 01MQ07018) and by the FP7-ICT program of the European Community, under the PASCAL2 Network of Excellence (ICT-216886).

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 335–342, 2011.

namely (1) it does not rely on a priori given similarities between tasks, but learns them via MKL, (2) it uses asymmetric task relations thus avoids negative transfer effects, i.e. good classifiers are not deteriorated by other bad classifiers, which may occur in MTL with symmetric task relations and (3) in contrast to other MTL methods it is scalable. Our experimental results on PASCAL VOC 2009 images show that information from the other classes can improve the classification performance significantly.

The rest of this paper is organized as follows. Section 2 describes related work. In Section 3 we explain MKL and our multi-task learning procedure. Experimental results are described in Section 4 and Section 5 concludes this work and discuss future issues.

2 Related Work

The main goal of multi-task learning is to achieve better performance by learning multiple tasks simultaneously. In general multi-task learning methods can either utilize common structure or use explicit relations between tasks. Methods utilizing common structure in the data can be used for combining multiple features or learning from unlabeled data in a multi-task framework [1,12]. Using relations between tasks became very popular in last years. For example Evgeniou *et al.* [3] proposed a framework in which relations between tasks are represented by a kernel matrix and multi-task learning is performed by using a tensor product of the feature and task kernel as input for SVM. Similar work can be also found in [9,11] or in [8] where the authors used Gaussian Processes for learning multiple-tasks simultaneously. All these approaches are theoretically attractive, but have drawbacks which reduce their applicability in practice. The dimensionality of the kernel matrix increases (as square) with the number of tasks, thus these methods are intractable for many real-world problems. Further, it is necessary to determine task similarities appropriately in advance and in contrast to our method these approaches assume a symmetric relationship between the tasks, but in practice a gain from task A on task B may incur a loss from task B on task A.

Our work is, in philosophy, close to Lampert and Blaschko [5] who applied MKL to multi-class object detection problems. However, their procedure cannot be used for object categorization where detection is not the primal interest and no bounding boxes of objects are available.

3 Multi-task Learning via MKL

3.1 Multiple Kernel Learning

In image categorization, combining many kernels $K_j(\mathbf{x}, \bar{\mathbf{x}})$, (similarity measures between images $\mathbf{x}$ and $\bar{\mathbf{x}}$) for $j = 1, \ldots, m$ constructed from various image descriptors has become a standard procedure. Multiple kernel learning (MKL) is a method which can choose the optimal weights $\{\beta_j\}_{j=1}^m$ of the combined kernel $\sum_{j=1}^m \beta_j K_j(\mathbf{x}, \bar{\mathbf{x}})$ and learn the parameters of support vector machine (SVM) simultaneously (see [6]).

Originally MKL imposes a 1-norm constraint $\|\beta\|_1 = \sum_j \beta_j = 1$ on the mixing coefficients to enforce *sparse* solutions. Recently, Kloft et al. [4] extended MKL to allow *non-sparse* mixing coefficients by employing a generalized p-norm constraint $\|\beta\|_p = \left(\sum_j \beta_j^p\right)^{1/p} \leq 1$ and showed that it outperforms the original one in practice.

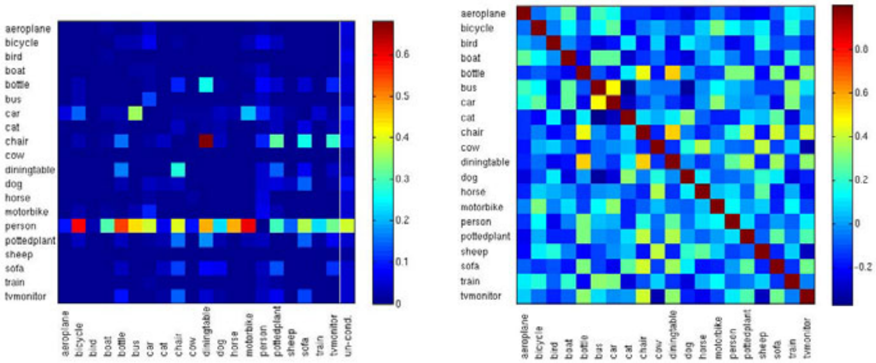


Fig. 1. Left Panel: Co-occurrence relations between the 20 categories of VOC 2009. The entries are $P(\text{class\_row}|\text{class\_column})$ except for the last column with $P(\text{class\_row})$. If a conditional probability is higher than its unconditioned value, then class_row appears frequently with class_column together (e.g. diningtable and chair). In the opposite case both categories are rather exclusive to each other (e.g. aeroplane and person). Right Panel: Kendall rank correlation scores.

3.2 Multi-task Learning

In a multi-task learning problem we obtain samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ where the vector $\mathbf{y}$ consists of the binary labels $y_c \in \{+1, -1\}$ of C different categories (e.g. visual concepts). Since some of these visual concepts may have semantic relations and can appear simultaneously in images, it is natural to incorporate such relations into learning processes. In order to see the pair-wise co-occurrence between the 20 object categories, we plotted in Figure 1 (left panel) the conditional probabilities of the classes in the rows given those in the columns, where the diagonal entries with probabilities 1 are excluded for better visibility. We see that for instance diningtable and chair appear together frequently, while classes such as cat and cow are rather exclusive to each other.

Apart from co-occurrence there exist other factors characterizing class-relationships. The right panel of Figure 1 shows Kendall rank correlation score τ

$$\tau = \frac{(\#\text{concordant pairs}) - (\#\text{disconcordant pairs})}{\frac{1}{2}n(n-1)},$$

where n is the number of samples and a pair of samples is concordant, if the orders of the two classification scores agree. For instance, although aeroplane and boat appear together very rarely, their classification scores are positively correlated, because they often share similar backgrounds (e.g. blue sky). Multi-task procedures aim at improving classification performances by uncovering statistical relations overlooked by class-wise binary solutions. However, in competitions on object categorization like PASCAL VOC 2009, there have been almost no submissions using the additional label interactions to improve performance over one-vs-rest classifiers.

We tackle this by constructing a classifier which incorporates information from the other categories via MKL for each class separately. Our procedure consists of two steps.

First Stage: For each binary problem we compute the SVM outputs using the average feature kernel.

Second Stage: (1) For each class we construct an output kernel based on the SVM scores from first stage. (2) For each class, we apply sparse or non-sparse MKL with the feature kernels and the output kernels from the other categories.

We measure similarities between the SVM outputs by using the exponential kernel

$$\widetilde{K}_c(\mathbf{x}_i, \mathbf{x}_j) = \exp\left[- \left\{s_c(\mathbf{x}_i) - s_c(\mathbf{x}_j)\right\}^2\right], \tag{1}$$

where $s_c(\mathbf{x}_i)$ is the score of SVM for the c-th category for the i-th example $\mathbf{x}_i$. We neither normalized the scores, nor optimized kernel width further. It must be noted that we cannot compare SVM outputs of training examples with those of test samples, because their statistical properties are not the same. In particular, most of the training images become support vectors whose SVM outputs almost coincide with their labels. In this paper, we deployed 5-fold cross-validation to obtain reasonable scores for the training images, while for the validation and test images, one SVM with the entire training data was used to compute the classifier outputs.

4 Experimental Results

4.1 Experimental Setup

We used the data set of PASCAL VOC 2009 [2] which consists of 13704 images and 20 object classes with an official split into 3473 training, 3581 validation, and 6650 test examples. We deployed the bag-of-words (BoW) image representations over various color channels and the spatial pyramid approach [7] using SIFT descriptors calculated on different color channels (see [10]). We used a vocabulary of size 4000 learned by k-means clustering and applied a χ^2 kernel which was normalized to unit variance in feature space. Average precision (AP) was used as performance measure.

We created 10 random splits of the unified training and validation sets into new smaller data sets containing 1764 training, 1763 validation, and 3527 test images. Since the true labels of the original test images have not been disclosed yet, we excluded these from our own splits. We remark that the AP scores reported in this paper are not comparable with those by the VOC 2009 winners, because the training sample size here is much smaller than that of the official split. For each of the 10 splits, the training images were used for learning classifiers, while with its validation part we selected the SVM regularization parameter C based on the AP score. The regularization constant C is optimized class-wise from the candidates $\{2^{-4}, 2^{-3}, \ldots, 2^{4}\}$.

The following three options were tested in the second step of our method: 1-norm MKL, 2-norm MKL and average kernel SVM with 15 BoW and 19 output kernels.

4.2 Performance Comparison

In the first experiment we compare the performance of the three multi-task learning strategies (denoted with ‘M’) using both the 15 feature and the 19 output kernels with

Table 1. Average AP results for 10 splits. The combination of feature and output kernels with non-sparse MKL outperforms all other settings. The performance gains of `M_2mkl` are all significant.

Split	`B_ave`	`B_2mkl`	`B_1mkl`	`M_ave`	`M_2mkl`	`M_1mkl`
1	0.4949	0.4932	0.4726	0.4958	**0.5033**	0.4737
2	0.4845	0.4845	0.4683	0.4818	**0.4926**	0.4675
3	0.4893	0.4882	0.4775	0.4879	**0.4945**	0.4756
4	0.4963	0.4957	0.4804	0.4938	**0.5004**	0.4804
5	0.4862	0.4910	0.4704	0.4910	**0.5000**	0.4781
6	0.4908	0.4896	0.4783	0.4929	**0.5029**	0.4779
7	0.4875	0.4905	0.4665	0.4962	**0.5012**	0.4685
8	0.4866	0.4875	0.4736	0.4857	**0.4970**	0.4753
9	0.4937	0.4959	0.4801	0.4980	**0.5067**	0.4852
10	0.4994	0.4983	0.4768	0.4887	**0.5030**	0.4788
average	0.4909	0.4914	0.4745	0.4912	**0.5002**	0.4761

three different baselines (denoted with '`B`') with kernels computed only from BoW features. The mean AP scores for the 10 splits are summarized in Table 1. Note that a comparison with standard MTL methods like [3] was not possible as it is computationally infeasible.

Two observations can be made from the results. First of all we see that the multi-task learning method with 2-norm MKL (`M_2mkl`) outperforms the other settings in all n = 10 runs. An application of the t-test shows that the performance gains are all highly significant e.g. the difference ΔX between `M_2mkl` and the average feature kernel baseline `B_ave` is significant with p-value less than 0.1% as

$$t = \sqrt{n}\frac{E[\Delta X]}{\sqrt{Var[\Delta X]}} = 7.4247$$

is larger than $t(1 - \frac{\alpha}{2}, n - 1)$ = 6.86.

The second observation which can be made from the results is that performance decreases when using 1-norm MKL. Note that this effect occurs among the '`B`' options as well which do not use any kernels from SVM outputs. This result is consistent with [4] who showed that the non-sparse MKL often outperforms 1-norm MKL in practice. Especially in difficult problems like object classification sparsity is often not the best choice as it ignores much information.

An interesting fact is that the performance gain is not uniformly distributed over the classes. The left panel of Figure 2 shows the average relative performance change between `M_2mkl` and `B_ave` over all 20 VOC classes. The largest average gain can be observed for the classes sheep, dog, diningtable, horse, motorbike and cow. Using the t-test we can show that the performance gain for the classes aeroplane, bicycle, bird, boat, car, cat, diningtable, dog, motorbike, person, sheep and train are significant with α = 5% and the gain for classes bird, boat, dog, motorbike, person and train is even significant with p-value less than 1%.

So the question now is *can we identify the classes (output kernels) which are responsible for the performance gain of these classes ?*

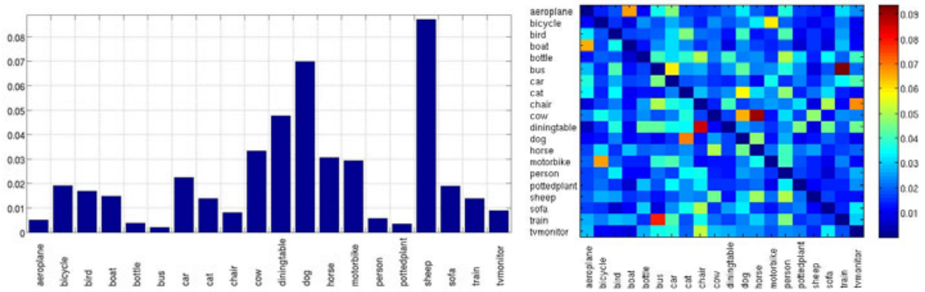


Fig. 2. Left Panel: Relative performance change (average over 10 runs) per class between the 2-norm multi-task MKL and average kernel baseline. Right Panel: Average MKL weights β. The classes in the row indicate the classification tasks, while those in the column are the output kernels. The strength of contribution is not symmetric e.g. see chair - diningtable.

4.3 Interactions between Object Categories

The kernel weights of MKL give a hint to what extent a classifier uses the output information from another classifier. The right panel of Figure 2 shows the average weights of MKL and we see some prominent links, e.g. train → bus, horse → cow, chair → diningtable. In order to visualize the relations, we created a class-relation graph with the 15 largest kernel weights β. This graph is asymmetric i.e. the output of the class A classifier may be important for classification of class B (arrow from A to B), but not vice versa. It is interesting that this graph although created from MKL kernel weights reveals a semantically meaningful grouping of the classes into: **Animals** (horse, cow, dog, cat), **Transportation** (bus, car, train), **Bikes** (bicycle, motorbike), **Home** (chair, tvmonitor, sofa, diningtable), **Big bluish areas** (aeroplane, boat).

We can think of at least two reasons why the classifier output can help classifiers in the same group. First, co-occurrence relationships can provide valuable information e.g. a chair in the image is an evidence for a diningtable. The second reason is that objects from different classes may have similar appearance or share similar context e.g. images with aeroplanes and boats often contain a large bluish area, the sky and water respectively, so that the outputs of aeroplane classifier may help to classify boats.

4.4 Ranking of Images

When we compare image rankings of the baseline and our classifiers in more detail, we gain interesting insights, e.g. the cat → dog and chair → diningtable rankings are analysed in Figure 4. On the horizontal axis we divide the test set into 35 groups of 100 images based on cat (or chair) scores and create a box plot of the rank difference of dog (resp. diningtable) outputs between the two procedures for each group. In both cases, our method provide better ranks (i.e. positive) in some interval from rank 101 (101 - 1300 for dog and 101 - 600 for diningtable). We conjecture that this is caused mainly by similarities between objects (e.g. cat and dog) and/or backgrounds (e.g. indoor scene). On the other hand, the top group (1 - 100) has rather negative shifts (i.e. our method gave lower ranks than the baseline) for dog, while shows more positive changes for diningtable. It can be possible that this behavior is caused by co-occurrence relations.

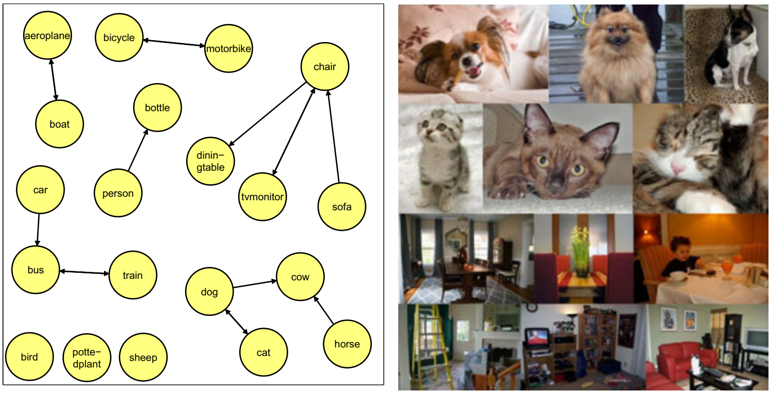


Fig. 3. Left Panel: Class-relation graph computed from MKL weights showing the 15 strongest relations. An arrow from A to B indicates that B is using the output kernel from A with high kernel weight. Right Panel: Images with substantial rank changes. Top images using dog classifier (upper: 297→207, 50→4, 140→60, lower: 33→164, 86→280, 108→1057), Bottom images using diningtable classifier (upper: 28→15, 486→345, 30→6, lower: 36→63, 35→61, 9→36).

Finally, we show in the right panel of Figure 3 example images which had large differences in rankings by the two methods. The three images in the upper row of each group got higher ranks by our classifier and contain the correct objects, while the other three in the lower row had worse ranks and the object class does not appear. For the images containing the correct objects, the proposed method gave better ranking than the baseline. Among dog (or diningtable) images, 63% (60% resp.) obtained better ranks

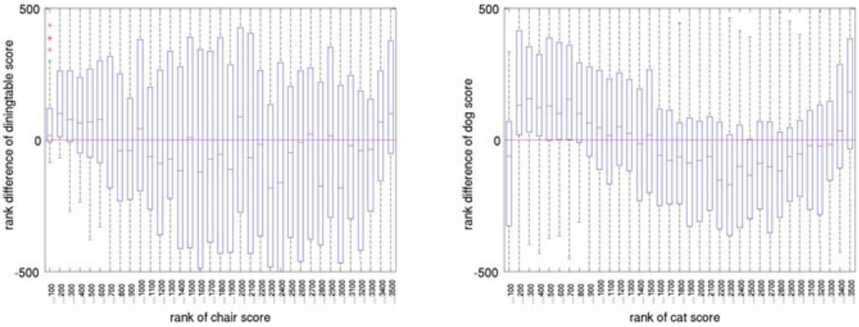


Fig. 4. Differences between the baseline and our method of diningtable (dog) ranks conditioned on chair (cat) ranks. On the horizontal axis each group consists of 100 images, i.e. the first group is from rank 1 to 100 of the chair (cat) classification score. The box plots show that till rank 600 the diningtable score tend to be ranked higher by our method than by the SVM baseline probably due to co-occurrence and common indoor context. In the top group of cat, we see a downside shift of the dog score probably because of negative co-occurrence relation, while images with rank 101-1300 of the cat score are ranked higher due to similarities between dog and cat images.

with median improvement +36 (+14 resp.). On the other hand, we also observed that mutually-exclusive categories may reduce false positives, e.g. among non-diningtable images containing cat, 73% had worse ranks with median difference -139.

5 Conclusions

The multi-task learning approach presented in this paper allows to automatically incorporate relations between object categories into the learning process without a priori given task similarities, by using the output information from other classifiers. It can potentially capture co-occurrence information as well as visual similarities and common backgrounds. We showed that our approach with non-sparse MKL not only significantly outperforms the baselines, but also allows to extract semantically meaningful relations between categories. Furthermore, we analysed the rankings obtained by the proposed method in comparison with those by the non-multi-task baseline. It reveals that interactions between different categories are affected by multiple factors, but can be captured by our method in a non-linear fashion through the output kernels.

In future research we plan to combine our idea with a boosting method and compare it with existing approaches which model relationships between classes explicitly.

References

1. Argyriou, A., Evgeniou, T., Pontil, M.: Multi-task feature learning. In: Neural Inf. Proc. Sys. MIT Press, Cambridge (2007)
2. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2009 (VOC 2009) Results
3. Evgeniou, T., Micchelli, C.A., Pontil, M.: Learning multiple tasks with kernel methods. J. of Mach. Learn. Res. 6, 615–637 (2005)
4. Kloft, M., Brefeld, U., Sonnenburg, S., Laskov, P., Müller, K.-R., Zien, A.: Efficient and accurate ℓ^p-norm mkl. In: Neural Inf. Proc. Sys. (2010)
5. Lampert, C., Blaschko, M.: A multiple kernel learning approach to joint multi-class object detection. In: Rigoll, G. (ed.) DAGM 2008. LNCS, vol. 5096, pp. 31–40. Springer, Heidelberg (2008)
6. Lanckriet, G.R.G., Cristianini, N., Bartlett, P., El Ghaoui, L., Jordan, M.I.: Learning the kernel matrix with semidefinite programming. J. of Mach. Learn. Res., 27–72 (2004)
7. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Conf. on Comp. Vision & Pat. Rec. (2006)
8. Ming, K., Chai, A., Williams, C.K.I., Klanke, S., Vijayakumar, S.: Multi-task gaussian process learning of robot inverse dynamics. In: Neural Inf. Proc. Sys. (2008)
9. Sheldon, D.: Graphical multi-task learning. In: Neural Inf. Proc. Sys. Workshop on Strictured Input - Structured Output (2008)
10. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE Trans. Pat. Anal. & Mach. Intel. (2010)
11. Wu, W., Li, H., Hu, Y., Jin, R.: Multi-task learning: Multiple kernels for multiple tasks. Technical report, Microsoft Research (2010)
12. Yuan, X.-T., Yan, S.: Visual classification with multi-task joint sparse representation. In: IEEE Conf. on Comp. Vision & Pat. Rec., pp. 3493–3500 (2010)

Multiple Random Subset-Kernel Learning

Kenji Nishida[1], Jun Fujiki[1], and Takio Kurita[2]

[1] Human Technology Research Institute,
National Institute of Advanced Industrial Science and Technology (AIST),
1-1-1 Umezono, Tsukuba Ibaraki, 305-8568, Japan
{kenji.nishida,jun-fujiki}@aist.go.jp
[2] Faculty of Engineering, Hiroshima University, Kagamiyama 1-7-1,
Higashi-Hiroshima Hiroshima 739-8521, Japan
tkurita@hiroshima-u.ac.jp

Abstract. In this paper, the multiple random subset-kernel learning (MRSKL) algorithm is proposed. In MRSKL, a subset of training samples is randomly selected for each kernel with randomly set parameters, and the kernels with optimal weights are combined for classification. A linear support vector machine (SVM) is adopted to determine the optimal kernel weights; therefore, MRSKL is based on a hierarchical SVM. MRSKL outperforms a single SVM even when using a small number of samples (200 to 400 out of 20,000 training samples), while the SVM requires more than 4,000 support vectors.

Keywords: Kernel Method, Multiple Kernel Learning, Support Vector Machine, Random Sampling.

1 Introduction

Recently, multiple kernel learning (MKL) has been proposed to improve the classification performance of single kernel classifiers [1,2]. Although the method based on the unweighted sum if multiple kernels is considered the simplest method, it may not be the ideal one.Therefore, various programming methods for finding the optimal combination weight have been proposed. Lanckreit [3] and Bach [1] proposed an efficient algorithm based on sequential minimal optimization (SMO).

The discriminant function for MKL is described as a weighted summation of kernel values:

$$f(x) = \sum_{m=1}^{p} \beta_m \langle \boldsymbol{w}_m, \Phi(\boldsymbol{x}) \rangle + b \tag{1}$$

where m indexes kernels. β_m is the weight coefficients for the kernel; $\boldsymbol{w}_m$, the weight coefficient for the sample; $\Phi_m(\boldsymbol{x})$, the mapping function for feature space m; and p, the number of kernels. Reforming equation (1) using the duality condition, we obtain

$$f(x) = \sum_{m=1}^{p} \beta_m \sum_{i=1}^{n} \alpha_{m_i} y_i \underbrace{\langle \Phi_m(\boldsymbol{x}), \Phi_m(\boldsymbol{x}_i) \rangle}_{\boldsymbol{K}_m(\boldsymbol{x}, \boldsymbol{x}_i)} + b \tag{2}$$

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 343–350, 2011.

where n is the number of sample; α_{m_i}, the weight coefficient; and y_i, be the sample label. The kernel weights satisfy the condition $\beta_m \geq 0$ and $\sum_{m=1}^{p} \beta_i = 1$. Different kernels (such as linear, polynomial, and Gaussian kernels) or kernels with different hyperparameters (for example, Gaussian kernels with different Gaussian widths) can be combined; however, the same weight is assigned to a kernel over all the input samples, as per the definition in equation (1).

Although in the original definition of MKL (equation (1)) different weights are not assigned to a kernel for different samples, kernels can be combined over different subsets of training samples, such as

$$f(x) = \sum_{m=1}^{p} \beta_m \sum_{i \in \dot{\boldsymbol{X}}} \alpha_{m_i} y_i \langle \boldsymbol{K}_m(\boldsymbol{x}, \boldsymbol{x}_i) \rangle + b \tag{3}$$

where $\dot{\boldsymbol{X}}$ stands for the subset of training samples for the mth kernel, while $\boldsymbol{X}$ stands for the full set of training samples. The sampling policy for the subsets is not restricted to any method, but if subsets are sampled according to the probability distribution $\eta_m(\boldsymbol{x})$, the kernel matrix is defined as follows:

$$K_\eta(\dot{\boldsymbol{x}}_i, \dot{\boldsymbol{x}}_j) = \sum_{m-1}^{p} \langle \Phi_m(\dot{\boldsymbol{x}}_i), \Phi_m(\dot{\boldsymbol{x}}_j) \rangle \tag{4}$$

where $\dot{\boldsymbol{X}} = \eta \boldsymbol{X}$. The probability that $K_\eta(\dot{\boldsymbol{x}}_i, \dot{\boldsymbol{x}}_j)$ is obtained becomes the product of the probabilities of obtaining $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Therefore, a subset kernel is determined by using the kernel matrix for all training samples and the sampling function η_m, as follows:

$$\begin{aligned} K_\eta(\dot{\boldsymbol{x}}_i, \dot{\boldsymbol{x}}_j) &= \sum_{m-1}^{p} \langle \Phi_m(\dot{\boldsymbol{x}}_i), \Phi_m(\dot{\boldsymbol{x}}_j) \rangle \\ &= \sum_{m=1}^{p} \eta_m(\boldsymbol{x}_i) \underbrace{\langle \Phi_m(\boldsymbol{x}_i), \Phi_m(\boldsymbol{x}_j) \rangle}_{\boldsymbol{K}_m(\boldsymbol{x}_i, \boldsymbol{x}_j)} \eta_m(\boldsymbol{x}_j), \end{aligned} \tag{5}$$

which is eventually equivalent to the definition of *localized multiple kernel learning* (LKML) [5].

For good classification performance in MKL, the optimal hyperparameters for the kernels and sample subsets (sampling function according to $\eta_m(\boldsymbol{x})$) must be determined by using different subsets; however, this requires an exhaustive search for the desired parameters and sampling functions. Therefore, we employ random sampling for the training subsets and randomly set hyperparameters for the kernels. The final classifier is determined by a linear combination of random kernels (randomly sampled subset and randomly set hyperparameters), and the β_m values are optimized to obtain the best classification performance.

In this paper, we propose *multiple random subset kernel learning* (MRSKL), a multiple kernel learning algorithm for a randomly selected subset of training samples. The proposed algorithm uses a small subset for each kernel, and the kernel values are combined according to the classification result obtained for all training samples. Simultaneous optimization of α_{m_i} and β_m in equation (3) has been a major interest in MKL research, as reported by Bach [1] and Rakotmamonjy [4], but the coefficients are independently optimized in the proposed algorithm.

The rest of the paper is organized as follows.

In Section 2, we describe the MRSKL algorithm. In Section 3, we present the experimental results for an artificial dataset. MRSKL showed good classification performance which exceeds the SVM result for the test samples.

2 Multiple Random Subset-Kernel Learning Algorithm

2.1 Learning Algorithms Using a Subset of Training Samples

Several algorithms that use a subset of training samples are proposed. These algorithms can be used to improve the generalization performance of classifiers or to reduce the computation cost for the training. *Feature vector selection* (FVS) [6] has been used to approximate the feature space F spanned by training samples by the subspace F_s spanned by selected *feature vectors* (FVs). *Import vector machine* (IVM) is built on the basis pf kernel logistic regression (KLR) and used to approximate kernel feature space by a smaller number of *import vectors* (IVs). While FVS and IVM involve approximation of the feature space by their selected samples, *RANSAC-SVM* [9] involves approximation of the classification boundary by randomly selected samples with optimal hyperparameters. In FVS and IVM, samples are selected sequentially, but in the case of RANSAC-SVM, samples are randomly selected; nevertheless, in all these cases, a single kernel function is used over all the samples.

SABI [8] sequentially selected a pair of samples at a time and carried out linear interpolation between the pair in order to determine a classification boundary. Although SABI does not use the kernel method, the combination of classification boundaries can be considered as a combination of different kernels.

An exhaustive search for the optimal sample subset requires a large computation; therefore, we employed random sampling to select subsets and combined multiple kernels with different hyperparameters for the subsets for MRSKL.

2.2 Subset Sampling and Training Procedure for MRSKL

Since the subset-kernel ($\boldsymbol{K}_m$) is determined by the subset of training samples (S_m), the subset selection strategy may affect the classification performance of each kernel. Therefore, in MKL using subset-kernels, the following three parameters must be optimized; sample weight α_{m_i}, kernel weight β_m, and sample subset S_m. However, since simultaneous optimization of three parameters is a very complicated process, we generate randomly selected subsets to determine α_{m_i}s for a subset kernel with randomly assigned hyperparameters; then, we determine β_m as the optimal weight for each kernel. When the kernel weights β_m are maintained to be optimal, the weights for kernels with insufficient performance becomes low. Therefore, such kernels may not affect the overall performance.

Separating the optimization procedures for α_{m_i} (sample weight) and β_m (kernel weight), we rewrite equation (2) by substituting $\alpha_{m_i} y_i \langle \boldsymbol{K}_m(\boldsymbol{x}, \boldsymbol{x}_i) \rangle$ with $f_m(x)$, as follows:

$$\begin{aligned} f(x) &= \sum_{m=1}^{p} \beta_m \sum_{i \in S_m} \alpha_{m_i} y_i \langle \boldsymbol{K}_m(\boldsymbol{x}, \boldsymbol{x}_i) \rangle + b \\ &= \sum_{m=1}^{p} \beta_m f_m(x) + b \end{aligned} \tag{6}$$

In MRSKL, we first optimize α_{m_i} for the subset-kernel classifier $f_m(x)$ and then optimize β_m.

The detailed MRSKL algorithm is as follows:

1. Let n be the number of training samples T; p, the number of kernels; and l, the number of samples in the selected subsets S_m,
2. Repeat the following steps p times
   (a) Determine Q training subsets S_m by randomly selecting samples from T
   (b) Randomly set hyperparameters (such as Gaussian width and regularization term for the RBF kernel)
   (c) Train the mth classifier f_m over the subset S_m
   (d) Predict all training samples T by f_m determining probability output
3. Train a linear SVM over f_m: $\{m = 1 \dots P\}$ to determine the optimal β_m for the final classifier

Parameter selection is performed by repeating steps 2b to step 2d, and the best parameter set is adopted in step 3.

RBF-SVM is employed for $f_m(x)$, and MRSKL is performed on the basis of a hierarchical SVM.

3 Experiment

The experimental results are discussed in this section. Although a wide variety of kernels are suited for use in MRSKL, we use only RBF-SVM for the subset-kernels to investigate the effect of random sampling. Hyperparameters (G and C for LIBSVM [10]) are randomly set to the desired range for the dataset. We employed linear-SVM to combine subset kernels to obtain the optimal kernel weight for classification.

3.1 Experimental Data

We evaluated MRSKL by using the artificial data in this experiment. The data are generated from a mixture of ten Gaussian distributions, five of which generate class 1 samples and others generate class –1 samples. 20,000 samples are generated for the training set, and 20,000 samples are independently generated for test set. The black contour in the figure 1 indicates Bayesian estimation of the class boundary; the classification ratio for

Bayesian estimation is 92.25% for the training set and 92.15% for the test set. The classification ratio for the full SVM, in which the parameters are determined by five-fold cross-validation ($c = 3$ and $g = 0.5$), is 92.22% for the training and 91.95% for the test set, with 4,257 support vectors.

The fitting performance of MRSKL may be affected by the subset selection policy; therefore, we first evaluated the performance by the smallest subset size, which includes one pair of samples from class 1 and class –1 each. All the experiments were run thrice, and the results were averaged.

3.2 A Single-Pair Subset-Kernel

Figure 1 shows the classification boundary in MRSKL for various numbers of kernels. From the result, a good classification boundary can be determined using as few as 100 samples (50 single-pair subset-kernels), while a larger number of samples would be required in an SVM.

Figure 2 shows the classification ratio for the training samples, and figure 3 shows the classification ratio for the test samples with the regularization parameter C for 2^{10} to 2^{-1} and the Gaussian width parameter G for 2^2 to 2^{-5}. The average classification ratio for the training samples became comparable to the SVM result for about 100 kernels but the classification ratio for the test samples exceeds the SVM result for 150 kernels. finally, the classification ratio reached 92.20% for 200 kernels. The average classification ratio for the test samples exceeded the SVM result for about 150 kernels and finally reached 91.97% for 200 kernels. Since each subset contained only one pair (two) of samples in this experiment, only 200 samples were required to attain a fitting performance similar ti that in the SVM case with 4,257 support vectors. This result for the training samples indicates that MRSKL can show high fitting performance with a

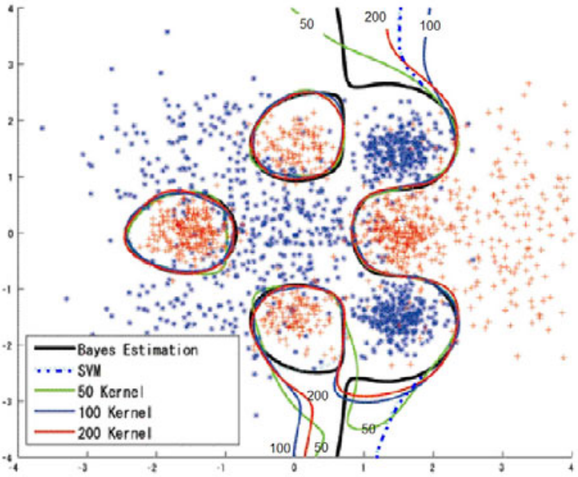


Fig. 1. MRSKL Classification Boundary with Single-Pair Kernel ($C = 2^{10}$ to 2^{-1}, $G = 2^2$ to 2^{-5})

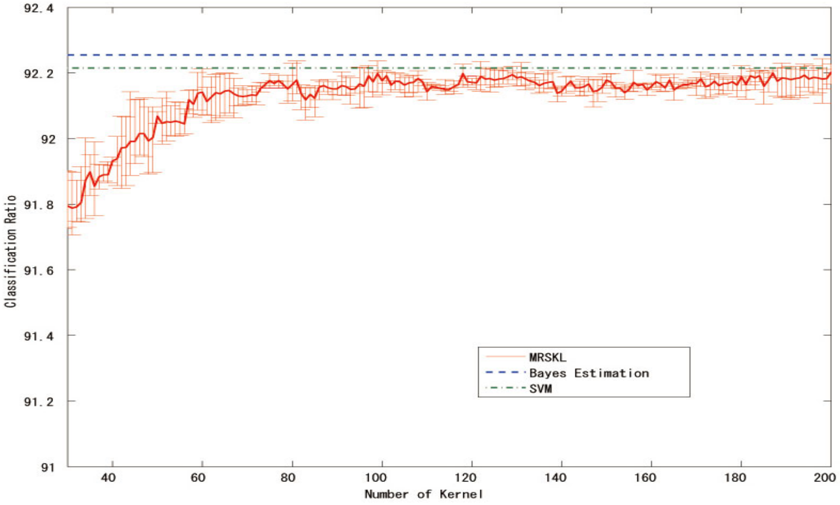


Fig. 2. MRSKL Result for Training Samples with Single-Pair Kernel ($C = 2^{10}$ to 2^{-1}, $G = 2^2$ to 2^{-5})

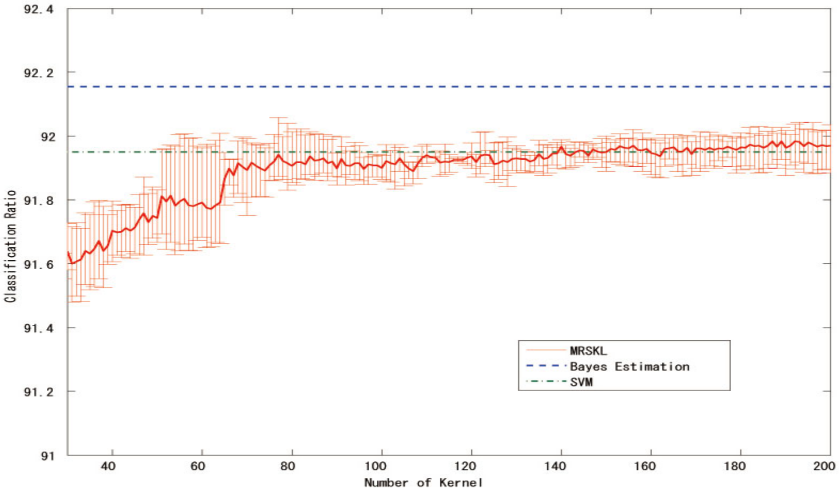


Fig. 3. MRSKL Result for Test Samples with Single-Pair Kernel ($C = 2^{10}$ to 2^-1, $G = 2^2$ to 2^{-5})

small number of support vectors than does an SVM. The result for the test samples indicates that MRSKL can show higher generalization performance than does an SVM.

3.3 Result for Benchmark Set

Next, we examined a benchmark set `cod-rna` from the LIBSVM dataset [11]. The `cod-rna` dataset has eight attributes 59,535 training samples, and 271,617 validation samples with two-class labels. Hyperparameters for a single SVM were obtained by performing grid search through five-fold cross-validation and randomly set for MRSKL around the values for a single SVM. We applied the random subset-kernel with parameter selection for this dataset, because the dataset includes a large number of samples. We examined 500-sample, 1000-sample, and 5000 sample subsets.

Table 1 shows the results for the `cod-rna` dataset. MRSKL outperformed the single SVM with a subset size of 1,000 (1.7% of the total number of the training samples) combining 2,000 kernels and with a subset of 5,000 (8.3% of the training samples) combining 100 kernels.

Table 1. Classification Ratio for `cod-rna` dataset

	Number of kernel	Training	Test
Single SVM (Full set)	1	95.12	96.23
MRSKL subset = 500	3000	95,03	96.16
MRSKL subset = 1000	2000	95.30	**96.30**
MRSKL subset = 5000	100	94.90	**96.24**

4 Conclusion

We proposed an MRSKL algorithm, which combines multiple kernels generated from small subsets of training samples.

The result for the smallest subset (one pair) showed that MRSKL could approximate the classification boundary with a small number of samples. The 200-pair (400-samples) subset-kernel outperformed the SVM with 4,257 support vectors.

A multiple-sample subset (100 samples) helped accelerate the convergence of the classifier, but the final classification performance for test samples showed only a small improvement.

The result for the benchmark dataset `cod-rna` showed that MRSKL with a subset size corresponding to 2% or 5% of the training samples can outperform the single SVM with optimal hyper-parameters.

Although 200 or 1000 kernels must be combined in MRSKL, the number of computations for the subset-kernels would not exceed that for a single (full-set) SVM, because an SVM requires at least $O(N^2)$ to $O(N^3)$ computations.

We employed a linear SVM to combine kernels and obtain the optimal kernel weights. However, this final SVM took up a majority of the computational time in MRSKL since it had to trained for as many samples as the large-attribute training samples.

In this study, we used all the outputs from subset-kernels for the training samples; however, we can apply feature selection and sample selection for the final linear SVM, as this may help reduce computation and improve the generalization performance simultaneously.

Acknowledgment. The authors would like to thank Prof. Toshikazu Wada of Wakayama University for providing us with valuable information on the relationship between prototyping and classification and also for drawing our attention to interesting papers.

This work was supported by JSPS KAKENHI 22500172.

References

1. Bach, F.R., Lanckriet, G.R.G., Jordan, M.I.: Multiple Kernel Learning, Conic Duality, and the SMO algorithm. In: Proc. International Conf. on Machine Learning, pp. 41–48 (2004)
2. Sonnnenburg, S., Röotsch, G., Schäfer, C., Schöolkopf, B.: Large Scale Multiple Kernel Learning. J. of Machine Learning Research 7, 1531–1565 (2006)
3. Lanckriet, G.R.G., Cristiani, N., Bartlett, P., Ghaoui, L.E., Jordan, M.I.: Learning the kernel matrix with semidefinite programming. J. of Machine Learning Research 5, 27–72 (2004)
4. Rakotomamonjy, A., Bach, F., Canu, S., Grandvalet, Y.: More Efficiency in Multiple Kernel Learning. In: Proc. of International Conf. on Machine Learning (2007)
5. Göonen, M., Alpaydin, E.: Localized Multiple Kernel Learning. In: Proc. if International Conf. on Machine Kearning (2008)
6. Baudat, G.: Feature vector selection and projection using kernels. NeuroComputing 55(1), 21–38 (2003)
7. Zhu, J., Hastie, T.: Kernel Logistic Regression and the Import Vector Machine. J. of Computational and Graphical Statistics 14(1), 185–205 (2005)
8. Oosugi, Y., Uehara, K.: Constructing a Minimal Instance-base by Storing Prototype Instances. J. of Information Processing 39(11), 2949–2959 (1998) (in Japanese)
9. Nishida, K., Kurita, T.: RANSAC-SVM for Large-Scale Datasets. In: Proc. ICPR 2008 (CD-ROM) (December 2008)
10. Chang, C.C., Lin, C.J.: A library for support vector machines (2001), `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
11. `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html#cod-rna`

Getting Robust Observation for Single Object Tracking: A Statistical Kernel-Based Approach

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. Mean shift-based algorithms perform well when the tracked object is in the vicinity of the current location. This cause any fast moving object especially when there is no overlapping region between the frames fails to be tracked. The aim of our algorithm is to offer robust kernel-based observation as an input to a single object tracking. We integrate kernel-based method with feature detectors and apply statical decision making. The foundation of the algorithm is patch matching where Epanechnikov kernel-based histogram is used to find the best patch. The patch is built based on Shi and Tomasi [1] corner detector where a vector descriptor is built at each detected corner. The patches are built at every matched points and the similarity between two histograms are modelled by Gaussian distribution. Two set of histograms are built based on RGB and HSV colour space where Neyman-Pearson method decides the best colour model. Diamond search configuration is applied to smooth out the patch position by applying maximum likelihood method. The works by Comaniciu et al. [2] is used as performance comparison. The results show that our algorithm performs better as we have no failure yet lesser average accuracy in tracking fast moving object.

Keywords: Tracking observation, Neyman-Pearson, Maximum likelihood, Patch matching.

1 Introduction

Getting the right observation for updating any tracking algorithm is a very challenging task. Tracking accuracy is highly dependent on good observation for getting a good performance. Thus, improving observation accuracy or observation association are the most crucial factors in building good tracker especially in people counting and behaviour analytics systems. In complex situations such as illumination change, clutter and occlusion; robust observations are hardly obtained which lead most trackers [3,2] resort to the prediction data alone. In certain situations, null observation or multiple false observations are no rare incidents which will hamper the tracking algorithms performance. In video analytics, the challenge of detecting the same object throughout the video is becoming more tedious as most of the objects are nonrigid where their appearance vary as the time goes on. Feature based algorithms such as SIFT [4] and SURF [5] will not give a good matching due to the fixed size kernel used for accumulating their

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 351–359, 2011.

descriptors which diminishes the accuracy when the point is occluded or blurred. On the other hand, histogram-based tracking algorithms [6,2] have been applied successfully for the nonrigid objects because of the matching is done based on the statistics of a group of pixels. The most popular histogram-based tracking is mean shift algorithm [7] where the next location is predicted based on the input of histogram backprojection via mean shift algortihm. Later Bradski introduced CAMSHIFT [6] which integrates scalability property into mean shift algorithm that allows the tracked object to have variable size. Kernel-based tracking which utilizes Epanechnikov kernel profile have been introduced by Comaniciu at el. [2]. This approach put more emphasize on the middle pixels and lesser weightage for further pixels during the accumulating the histogram's values. They also applied Bhattacharyya distance [8] for comparing between the two histograms. Another good histogram matching algorithms are such as the works by Funt and Finlayson [9] and earth mover distance [10], where both approaches are robust to illumination change.

In general, kernel-based algorithm performs well for single object tracking. However, as the scene become crowded and more objects need to be tracked, the algorithms start to falter, especially during occlusion. The works by Namboodiri et al. [11] and Peng et al. [12] are attempted to solve the problem of occlusion. The algorithm of Namboodiri et al. tweaks the localization of the mean shift by applying both forward and reverse methods so that it converges to the true modes. They also add scalability by utilizing SIFT's scale. However this contradicts their main argument where their algorithm should work in real time as it is a known fact that SIFT takes more processing power compared to the mean shift methods. Another approach by Peng et al. focuses on improving the updating method for the object model where Kalman filter is used. The predicted object model is called candidate model while the previous model is called current model and hypothesis testing is applied to choose the right histogram model. In the paper by Leichter et al. [13], they improve kernel-based method by using multiple object model so that it tracks well under sudden view change. This methods required the user to initialize the object model in several views which can be quite problematic. The main weakness of the mean shift algorithm is it depends on the proximity property. The tracker is prone to failure when the objects movement is fast. The algorithm of Li et al. [14] approaches this problem by extending the search area based on their hexagon method. However, it is a brute force search which requires significant processing cost. Besides, the algorithm still fail if the object moves fast enough to be outside of their search region.

The main goal of this paper is to offer a robust method of obtaining a good tracking observation. Firstly, the object bounding box in the initial frame will be the object model reference and several patches are built in the next frame as the target models. This is done by matching points of interest between two consecutive frames for building possible patches at matched vector location. This allows our method to search the tracked object for the whole image selectively. At each patch, their histograms are obtained by applying Epanechnikov kernel [13]. The matching between object model and target models are done by modelling the histograms with Gaussian distribution. Maximum likelihood method is applied to find the most probable patch. Two colour spaces are used; RGB and HSV for solving illumination change problem. RGB is good in normal condition while hue performs better under illumination change. Neyman Pearson method is applied to find the right colour model. Then, diamond search configuration is used in building position smoothing patches where once again maximum likelihood is applied.

2 Statistical Kernel-Based Observation (SKBO)

Basically, our algorithm (SKBO) has a close resemblance to the content-based image retrieval approach because of the matching procedure, but in this case the database is generated in a second image frame instead of via a predefined database. The output of the algorithm then can be applied as an input to filtering algorithm. The main four components of the algorithm are:

1. Generate points of interest.
2. Generate possible patches.
3. Undertake patch matching.
4. Perform position smoothing.

2.1 Generate Points of Interest

The purpose of finding points of interest is to generate early locations of where the possible patches may be built. At every located points, vector descriptor is built which will be compared between consecutive frames (F^t and F^{t-1}) to find the possible anchor point where the patches are built. Initially, the user will define the bounding box of the interest object. The importance of this user defined patch is that it serves as the reference for building the statistical data used in matching and smoothing procedures and in particular the reference histograms. Moreover, the size of the first frame patch indicates the original object size. Let P_w and P_h denote the width and height of the user defined bounding box, while (**x**)=(x, y) represents the coordinate location and t is the timing of the frame. Corner detectors as defined by Shi and Tomasi [1] are applied to find the possible points of interest. The threshold used for the Shi and Tomasi algorithm is around 0.01 which signifies the minimum eigenvalue required for the point to be considered as a corner. The minimum distance between consecutive corners is set at 3 pixels. This corner detector was chosen because of its ability to generate the points even during low ambient illumination and for low textured objects.

2.2 Generate Possible Patches

Possible patches, (β) are generated at each corner where the vector descriptor, **V** are matched. There are 3 sets of vectors for each point of interest which are corresponding to red, green and blue channels, ($\mathbf{V}_R^{x,y,t}$, $\mathbf{V}_G^{x,y,t}$, $\mathbf{V}_B^{x,y,t}$). The advantage of this approach is we can expand the search area all over the frame while keeping the computational burden low as the histograms are built at selected points only. Contrary to mean shift approach where the processing cost is directly proportional to search area region. The vectors are generated by finding the colour difference between the anchor pixel and its selected neighbourhood pixels as shown in Figure 1. Let i denote the channel type and define

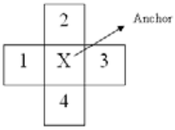


Fig. 1. Neighbourhoods pattern used for vectors generation

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

Then the vector components are sorted from the lowest to the highest value. Sorting allows the algorithm to find good match even during illumination change at the expense of orientation accuracy. This will not affects the algorithm performance as the objective is to built as many as possible good patch candidates. The decision rule (dr) for matching the vectors 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. Let $L_1^{x,y,t}$ denotes the label which takes the value 1 when the vectors are matched and 0 for unmatched vectors.

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

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

Each of the matched vector ($L_1^{x,y,t} = 1$) is candidate for locations at which patches are built. Patches for the second frame are anchored at the corner of the matched vector. Figure 2 shows an example of how the bounding box is generated. A subsequent test for distinguishing overlapping patches is performed after all patches have been assigned their location and size. This is done in order to reduce the calculation burden by reducing the number of patches. Moreover, most of the small differences occur because of "noise" in the patch generation process. The decision rule, L_2 for determining overlapping patches is calculated as in equation 4. Patch smoothing is performed if the overlapping area is more than $\gamma_2\%$ of the original patch size.

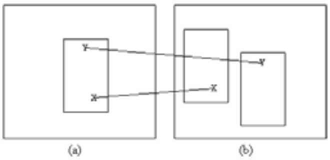


Fig. 2. Examples of constructing the new patches between the frames. The bounding boxes are aligned with respect to the matched vectors in the first frame. (a) First frame (b) Second frame.

$$L_2 = \begin{cases} 1 \text{ if the overlap region } > (\gamma_2).P_w.P_h \\ 0 \text{ if the overlap region } \leq (\gamma_2).P_w.P_h \end{cases} \tag{4}$$

The new combined patch location, $(\bar{x}, \bar{y})$ is the average of the corresponding center of the overlapping patches where N is the number of L_2s detected.

$$(\bar{x}, \bar{y}) = \left(\frac{1}{N} \sum_{j=1}^{N} x_j, \frac{1}{N} \sum_{j=1}^{N} y_j \right), \ (x_j, y_j) = j^{th} \text{ patch with } L_2 = 1 \tag{5}$$

2.3 Patch Matching

In order to choose which patch is the most probable one, we apply maximum likelihood method. The basis for our patch matching is by utilizing the histogram similarity. All histograms are built by applying Epanechnikov kernel, ($\mathcal{K}(\mathbf{x})$) [13] where more weightage is emphasize for the middle part and lesser as the pixels are further away from the anchor pixel. This is based on the assumptions that outliers and "noise" are more apparent at the kernel border. Before applying the kernel, each patch size is normalized. The kernel's profile is

$$\mathcal{K}(\mathbf{x}) \propto \begin{cases} (1 - \mathbf{x}) & \text{if } \ 0 \leq \mathbf{x} \leq 1 \\ 0 & \text{if } \ \mathbf{x} > 1 \end{cases} \tag{6}$$

Two colour models are applied consecutively, RGB and HSV. For RGB colour space, a 3-dimensional histogram is built for each patch while for HSV colour space, a 1-dimensional histogram is built based on hue component only. Under normal circumstances, RGB colour space works the best while under illumination change, hue channel is found to be invariant, yet the distinctive property is degraded. Histogram matching is done by modelling the relationship between two histograms as a Gaussian distribution. We find that Gaussian approach is good enough to model the histograms difference and directly applicable to maximum likelihood method. We assume the variances (σ) are equivalent for all channels and the covariance matrix is a diagonal matrix for simplicity purpose. Let $\mathbf{n}$ and $\mathbf{m}$ represent the histograms and i is the number of bins in 1-dimension. For 1-dimensional histogram, they are $n_{i\times 1}$ and $m_{i\times 1}$ elements while for 3-dimensional histogram, they are $n_{i\times 3}$ and $m_{i\times 3}$ elements.

$$P(\mathbf{n}; \mathbf{m}, \sigma^2) = \frac{1}{(2\phi)^{k/2}|\sigma|} \exp\left(-\frac{1}{2\sigma} (\mathbf{n} - \mathbf{m})^2 \right), k = \text{histogram dimension} \tag{7}$$

Every histograms are normalized before comparison is made. For each colour space, we use maximum likelihood approach to find the best match. The likelihoods are modelled by equation 7, $P(\mathbf{Y}|\beta) = P(\mathbf{n}; \mathbf{m}, \sigma^2)$ where $\hat{\beta}$ denotes the matched patch, $\mathbf{Y}$ represents the observation and j can be either RGB or hue.

$$\hat{\beta}_j = \underset{\forall \beta j}{\text{argmax}} \, P(\mathbf{Y}|\beta_j) \tag{8}$$

There are two candidates for the most likely patch, one is the output of RGB space while the other one is from Hue component. Neyman-Pearson hypothesis testing is applied to choose the most likely patch. Let $P(\mathbf{x};\mathcal{H}_0)$ be $P(\hat{\beta}_{\text{RGB}})$, $P(\mathbf{x};\mathcal{H}_1)$ be $P(\hat{\beta}_{\text{hue}})$ 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 and $\hat{\beta}_{\text{RGB}}$ is selected, while if $\mathcal{H}_1$ is chosen, ϵ is equal to 0 and $\hat{\beta}_{\text{hue}}$ is selected. In the next section, we will utilize only one colour space depending on the parameter $epsilon$. Figure 3 shows an example of selecting the right patch between frames.

$$\text{NP} = \frac{P(\mathbf{Y};\mathcal{H}_1)}{P(\mathbf{Y};\mathcal{H}_0)} = \frac{P(\hat{\beta}_{\text{hue}})}{P(\hat{\beta}_{\text{RGB}})} > \lambda_1 \tag{9}$$

$$\beta_{\text{fin}_1} = \begin{cases} \hat{\beta}_{\text{RGB}} \text{ if } P(\hat{\beta}_{\text{hue}}) < \lambda_1 P(\hat{\beta}_{\text{RGB}}) \\ \hat{\beta}_{\text{hue}} \text{ if } P(\hat{\beta}_{\text{hue}}) \geq \lambda_1 P(\hat{\beta}_{\text{RGB}}) \end{cases} \tag{10}$$

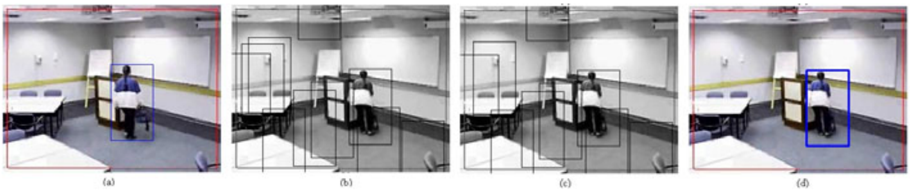


Fig. 3. Procedures for selecting the right patch. (a) Original patch (b) Raw patches (c) Combined patches (d) Maximum correlation patch

2.4 Position Smoothing

Usually the output from the patch matching is not aligned nicely with the tracked object. This error is prevalent during illumination change or in low ambient illumination. To smooth out the position, we apply once again maximum likelihood method as in equation 7. The translation test for adjusting the patch location is performed towards 4 directions: 1) leftward ($\beta_{\text{fin}_1}^{\text{new}_1}$), 2) upward ($\beta_{\text{fin}_1}^{\text{new}_2}$), 3) rightward ($\beta_{\text{fin}_1}^{\text{new}_3}$), and 4) downward ($\beta_{\text{fin}_1}^{\text{new}_4}$) as shown in Figure 5. The solid patch is the original position while the dashed patch is the moved patch by a step size value. The histograms of each of the

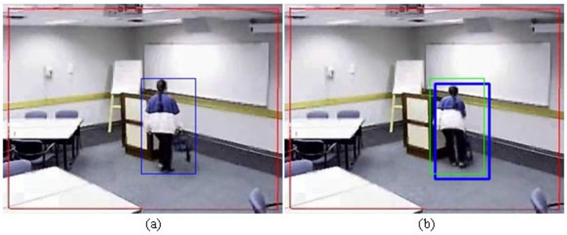


Fig. 4. Example of applying location smoothing algorithm. (a) Original patch (b) New location is indicated by the green bounding box

Figure 4 shows the example of applying position smoothing algorithm to fit the tracked object nicely. Diamond search method [14] is used to generate the possible shift patches anchored around β_{fin_1}. The colour space used is decided based on the ϵ value. Firstly, the step size (δ) used for adjusting the patch translation is determined as follows, $\delta = 0.1(\min(P_w, P_h))$.

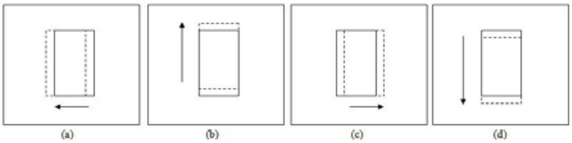


Fig. 5. Patches coordination for location smoothing (a) Left side translation (b) Upward translation (c) Right side translation (d) Downward translation

five patches including the original position patch are obtained. The likelihood of each patch is calculated with respect to the object model histogram. Let $\hat{\beta}_{\text{fin}_2}$ denotes the output of the position smoothing, $\beta_{\text{fin}_2} = \underset{\forall \beta_{\text{fin}_1}}{\operatorname{argmax}} P(\mathbf{x}|\beta_{\text{fin}_1})$. 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 the new β_{fin_2}. The algorithm is iterated until the estimated patch position remain constant as shown by L_3. This final patch then is ready to be fed into filter-based tracker such as Kalman as the measurement input.

$$L_3 = \begin{cases} \beta_{\text{fin}_2} = \beta_{\text{fin}_1}^{\text{old}_0} \text{ stop the iteration} \\ \beta_{\text{fin}_2} \neq \beta_{\text{fin}_1}^{\text{old}_0} \text{ continue the iteration} \end{cases} \tag{11}$$

3 Results and Discussion

The algorithm has been tested on several videos that contained fast moving object and sudden illumination change. We compared our algorithm with the kernel-based tracking by Comaniciu et al. [2] with slightly alteration where we use HSV colour space so that the comparison is fairer since the videos contain a lot of illumination change scenes. All videos have a size of 320 × 240 and samples of the videos are shown in the Figure 6. The parameters used are $\sigma = 122$, $\gamma_1 = 12$ and $\gamma_2 = 0.7$. Figure 6(a) to 6(c) contain a fast moving object while Figure 6(d) contains an object under sudden illumination change. We use relative error mean, $\mathcal{E}$ and number of failed tracks to quantify the algorithms performance. Relative error are the ratio between Euclidean distance of the object compared to the ground truth over its dimension. Let σ^{sim} denotes the simulation result and σ^{gnd} represents the ground truth.

$$\mathcal{E} = \left(\frac{\sqrt{(\sigma_x^{\text{sim}} - \sigma_x^{\text{gnd}})^2 + (\sigma_y^{\text{sim}} - \sigma_y^{\text{gnd}})^2}}{\max(P_w, P_h)} \right) \tag{12}$$

Table 1 shows that after averaging, our algorithm works better for video 1, while worse for video 3 and is equally good for video 2 compared to the kernel-based method. However, we manage to track the object at every frames while kernel-based method failed occasionally and needs to be reinitialized. Our lack of accuracy in video 3 is caused by the small number of feature points are generated. For video 4, we have a one

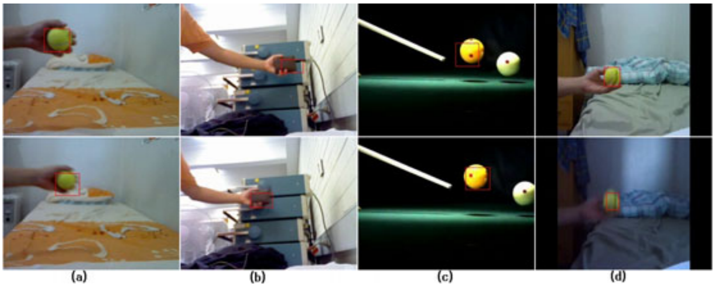


Fig. 6. Samples of the tracked object. (a) video 1 (b) video 2 (c) video 3 (d) video 4

Table 1. Performance evaluation of SKBO

Input video	Relative error mean		No. of tracking failure	
	Comaniciu et al.	SKBO	Comaniciu et al.	SKBO
video 1	0.5807	0.2251	4	0
video 2	0.3080	0.3144	2	0
video 3	0.1710	0.2638	1	0
video 4	0.1974	0.3850	0	1

failure out of several illumination changes test which leads to higher average relative error. Besides, we apply HSV colour model for the kernel-based method which help them improve the algorithm due to the hue invariant under lighting change.

4 Conclusion

In conclusion, we proved that our method works better in obtaining observation for single object tracking for fast moving object. The novelty of the approach is the search region is the entire frame but histograms are built at the matched points only. This method allows us reduce the computational cost significantly instead of brute force search. Besides, we also apply kernel profile in building the histograms and apply Gaussian based modelling for quantifying the histograms similarity. However, the performance is lower than kernel-based method under illumination change but still yield a reasonable accuracy. The performance can be improved if better feature detector is used but it will cost more computational power which hinders online applications.

References

1. Shi, J., Tomasi, C.: Good features to track. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 593–600 (1994)
2. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(5), 564–577 (2003)
3. Wren, C., Azarbayejani, A., Darrell, T., Pentland, A.: Pfinder: real-time tracking of the human body. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 780–785 (1997)

4. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157 (1999)
5. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (surf). Computer Vision and Image Understanding 110, 346–359 (2008)
6. Bradski, G.R.: Computer vision face tracking for use in a perceptual user interface. Intel. Technology Journal Q2, 705–720 (1998)
7. Fukunaga, K., Hostetler, L.D.: The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on Information Theory 21, 32–40 (1975)
8. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions. Bulletin of the Calcutta Mathematical Society 35, 99–109 (1943)
9. Funt, B.V., Finlayson, G.D.: Color constant color indexing. IEEE Transactions on Pattern Analysis and Machine Intelligence 17(5), 522–529 (1995)
10. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision 40, 99–121 (2000)
11. Namboodiri, V.P., Ghorawat, A., Chaudhuri, S.: Improved Kernel-Based Object Tracking Under Occluded Scenarios. In: Kalra, P.K., Peleg, S. (eds.) ICVGIP 2006. LNCS, vol. 4338, pp. 504–515. Springer, Heidelberg (2006)
12. Peng, N.S., Yang, J., Liu, Z.: Mean shift blob tracking with kernel histogram filtering and hypothesis testing. Pattern Recognition Letters 26(5), 605–614 (2005)
13. Leichter, I., Lindenbaum, M., Rivlin, E.: Mean shift tracking with multiple reference color histograms. Computer Vision and Image Understanding 114, 400–408 (2009)
14. Li, Z., Xu, C., Li, Y.: Robust object tracking using mean shift and fast motion estimation. In: International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2007, pp. 734–737 (2007)

Visual Words on Baggage X-Ray Images

Muhammet Baştan, Mohammad Reza Yousefi, and Thomas M. Breuel

Image Understanding and Pattern Recognition Group,
Technical University of Kaiserslautern,
Kaiserslautern, Germany
mubastan@gmail.com,yousefi@iupr.com,tmb@cs.uni-kl.de

Abstract. X-ray inspection systems play a crucial role in security checkpoints, especially at the airports. Automatic analysis of X-ray images is desirable for reducing the workload of the screeners, increasing the inspection speed and for privacy concerns. X-ray images are quite different from visible spectrum images in terms of signal content, noise and clutter. This different type of data has not been sufficiently explored by computer vision researchers, due probably to the unavailability of such data. In this paper, we investigate the applicability of bag of visual words (BoW) methods to the classification and retrieval of X-ray images. We present the results of extensive experiments using different local feature detectors and descriptors. We conclude that although the straightforward application of BoW on X-ray images does not perform as well as it does on regular images, the performance can be significantly improved by utilizing the extra information available in X-ray images.

Keywords: X-ray image analysis, bag of visual words, classification, retrieval.

1 Introduction

X-ray imaging is an important technology for security applications. Traditionally, images recorded by X-ray machines at security checkpoints are monitored by specially-trained screeners to prevent the entrance of illicit materials/objects (e.g., explosives, guns). Considering the huge number of luggages to be inspected (especially at the airports), it is desirable to (semi-)automate the inspection process to reduce the workload of screeners and hence increase the inspection speed and operator alertness. Automated inspection is also desirable for privacy concerns.

Our aim is to reduce the workload of X-ray screeners at airports by automatically analyzing and recognizing illicit materials/objects in baggage X-ray images such that the screeners will only need to check a small subset of the baggages that will be marked as suspicious by our automatic system. This is relevant to object detection, recognition and retrieval in computer vision, and has been studied extensively with some success on regular, visible spectrum images. However, X-ray images are quite different from the visible spectrum images: (1) they are transparent, pixel values represent the attenuation by multiple objects; (2) they may be very cluttered; (3) they are noisy due to the low energy X-ray imaging (Figure 1).

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 360–368, 2011.

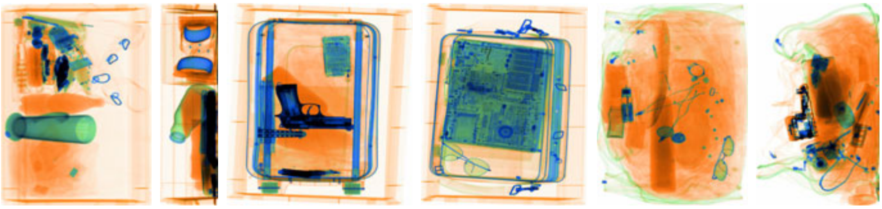


Fig. 1. Colored X-ray images

Objects in visible spectrum images are opaque and occlude each other. On the contrary, X-rays penetrate the objects, therefore, the objects along the X-ray path attenuate the signal and affect the final intensity value. Due to this, pixel intensities in X-ray images represent signal attenuation due to (multiple) objects. Contrast between objects in X-ray images is due to the differential attenuation of the X-rays as they pass through the objects. Attenuation of X-rays as they travel through objects is formulated by

$$I_x = I_0 e^{-\mu x} \tag{1}$$

where I_x is the intensity of the X-ray at a distance x from the source, I_0 is the intensity of the incident X-ray beam, and μ is the linear attenuation coefficient of the object measured in cm^{-1}. The higher the value of μ, the higher the attenuation.

Dual energy X-ray imaging combines two radiographs acquired at two different energy levels, to obtain both the density and atomic number of the materials, thus to provide information about material composition or at least improve image contrast [13,6]. The low-energy and high-energy images are fused with the help of a look-up table into a single color image to facilitate the interpretation of the baggage contents (Figure 2). Most X-ray machines record images from multiple view points (e.g., from 4 different viewing angles) so that the contents of the baggage can be seen better. This means, multiple images are recorded for each baggage, each image from a different view.

This different type of data has been left unexplored (except for a few works mentioned below), probably due to the unavailability of such data to computer vision researchers. We are investigating the applicability, and if possible adaptation and extension, of relevant computer vision algorithms on baggage X-ray data for the recognition and retrieval of baggages containing illicit materials/objects. This paper presents the results of extensive experiments on the classification and retrieval of baggage X-ray images using the popular bag of visual words method. We conclude that the classification and retrieval performance can be significantly improved by utilizing the extra information available in X-ray images, such as the dual energy images, multiple views and material-specific colors.

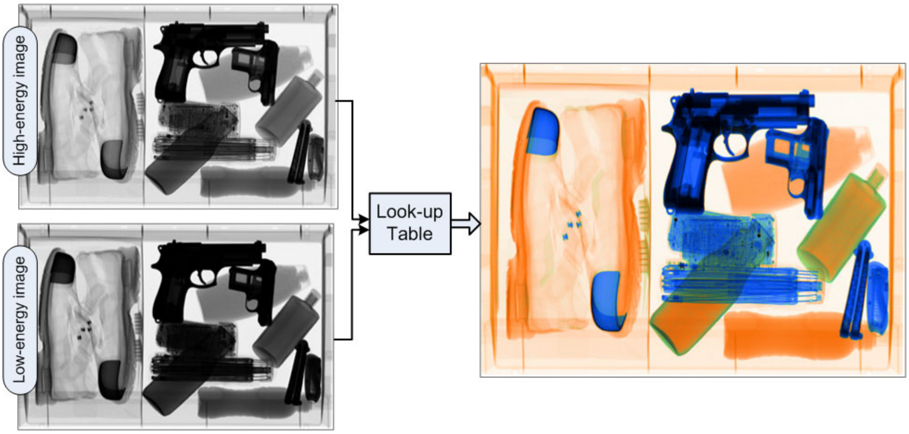


Fig. 2. Using two energy levels to obtain a color X-ray image using a look-up table. Colors represent the type of materials (atomic number); e.g., orange: organic materials, blue: non-organic materials (e.g., metals). Intensity is related to the thickness of the materials.

2 Related Work

The literature on baggage X-ray image analysis is very limited. There are only a few previous works, which mainly focused on image enhancement and segmentation. Chen *et. al* developed a combinational method to fuse, de-noise and enhance dual-energy X-ray images for improved threat detection and object classification [6]. A similar approach is presented in [9] for the enhancement of dual-energy X-ray carry-on luggage images. Singh and Singh proposed an approach for the optimization of the image enhancement tools [14]. Abidi *et. al* proposed an enhancement method for data decluttering for low density and hard-to-distinguish objects in baggage X-ray images.

Ding *et. al* proposed an attribute relational graph (ARG) based segmentation method for the segmentation of overlapping objects in X-ray images [8]. Similarly, Wang *et. al* proposed a structural segmentation method based on ARG matching [16]. Heitz *et. al* proposed a method for separating objects in a set of X-ray images using the property of additivity in log space, where the log attenuation at a pixel is the sum of the log-attenuations of all objects that the corresponding X-ray passes through [10].

To the best of our knowledge, there is no published work addressing the problem of recognition and retrieval on baggage X-ray images. Recently, Bag of (Visual) Words (BoVW or BoW) approaches have been successfully used in image classification, object detection/recognition and retrieval [7,15,12,2]. Adapted from text processing domain, BoW relies on visual words, which are generated by clustering a large number of local features obtained by computing low-level descriptors around local image patches. Described next is how we apply the BoW method on baggage X-ray images.

3 Method

We applied the standard BoW method for the classification and retrieval of X-ray images, and then utilized the extra information available in baggage X-ray images to improve the performance. The method consists of four main stages: (1) detection and description of image patches, (2) clustering the descriptors for vocabulary construction, (3) computing the BoW representation of images, and (4) classification and retrieval using the BoW representation [7].

Detection and description of image patches. Local image descriptors have proved to be successful in image classification, recognition and retrieval, since they are robust to partial occlusion and clutter, and image variations like translation, rotation and scaling. Several techniques have been suggested for the detection of local image patches [12]:

- Sparse representation with interest points
- Multiple interest point operators
- Dense sampling over a regular grid
- Random sampling

We used the first three sampling strategies in this work and experimented with several interest point detectors: DoG, Hessian-Laplace, Harris, FAST and STAR, which are all available in OpenCV [1]. Once the patches/points are detected, a descriptor can be computed around each patch/point. We experimented with three different descriptors: SIFT [11], SURF [3] and BRIEF [4].

Visual vocabulary construction. The visual vocabulary is obtained by clustering the visual descriptors into a number of clusters (e.g., 500, 1000, 2000) and taking the cluster centroids as the visual words. We used the well-known k-means clustering algorithm and also experimented with the Self Organizing Maps (SOM).

Vector quantization and BoW computation. After obtaining the visual vocabulary, images or image regions can be described by a histogram of visual words, the so-called bag of visual words. The histogram can be constructed using different assignment/weighting techniques, which affect the performance significantly [17].

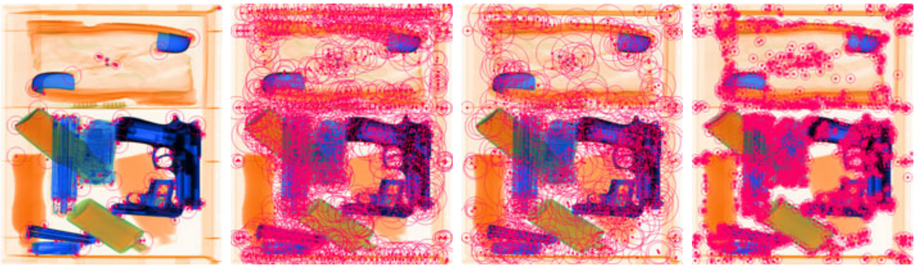


Fig. 3. Interest point detectors on a color X-ray image using the OpenCV 2.2 implementations of Harris, SIFT's DoG, SURF's Hessian-Laplace and FAST respectively

- Hard assignment (HA): assign each descriptor to the single closest visual word
- Soft weighting (SW): assign some weight to closest K visual words (e.g., K=3)
  - SW1: assign constant weights to closest K visual words (e.g., for K=3, [0.5, 0.25, 0.25])
  - SW2: determine the weights according to the distances to closest K visual words

Classification and retrieval. The bag of visual words representations of images can be used for supervised classification and retrieval. Among many available supervised classifiers, the Support Vector Machines (SVM) have been the most popular due to their performance [7,2]. For class-specific retrieval (searching for an instance of a class, e.g, handgun in X-ray images), the classification scores are used to rank the images.

4 Experiments

In this section, we present experimental results for the classification and retrieval of baggages containing *handguns* (Figures 1, 3, 4). Our dataset consists of baggage X-ray images recorded on a dual-energy X-ray machine, which provides 4 views (one from top, one from side and two views at some angle) for each baggage, hence a total of 12 images (1 low-energy, 1 high-energy and 1 color image for each of the 4 views) for each baggage. The imaged baggages are packed with a variety of objects/materials such as clothes, shoes, bottles, mobile phones, laptops, umbrellas, guns and knives. In practice, the baggages may (and does) contain anything that can fit in the baggage. The image sizes depend on the size of the baggages, and are around 600 × 700 pixels.

We used OpenCV 2.2 [1] for the interest point detectors and descriptors, and LIBSVM [5] with several kernels (linear, RBF, chi-square, intersection) for classification. We observed that the intersection kernel performed the best. Therefore, all the results presented below were obtained using the intersection kernel.

The training set contains 208 images (corresponding to 52 baggages), 52 positive images (13 baggages containing handguns), 156 negative images (39 baggages not containing handguns). The test set contains 764 images (191 baggages), 40 positive (10 baggages) and 724 negative images (181 baggages). We used all of the images from all 4 views in training, as if they were standalone images, without considering if they belong to the same baggage or not.

For performance evaluation, we used the standard recall, precision and average precision (AveP) measures. These performance measures were computed for both image- and baggage-based classification and retrieval. In the baggage-based evaluation, if an image is classified as positive, the baggage it belongs to is classified as positive. Similarly, for baggage-based retrieval, a baggage's score is taken as the highest score of all the images corresponding to the four views and ranking is done accordingly. This is because one of the four views of a baggage may contain a better view of the object than the other views, and hence it scores high in classification.

4.1 Results

Due to space limitations we present only the most relevant results. All the results were obtained using k-means for vocabulary learning (SOM performed worse), soft weighting (SW2, which worked best) in BoW computation, and histogram intersection kernel in classification with LIBSVM. We observed that DoG, Hessian-Laplace, Harris and FAST perform competitively, while STAR keypoints and dense sampling are not as good. Moreover, SIFT performed the best among the three descriptors.

Table 1 shows the recall, precision and average precision values for classification and retrieval for 3 different images for each view of a baggage with DoG detector and SIFT descriptor. Compared to visible spectrum images, the recall and precision values are considerably lower, indicating the difficulty of the X-ray image data. The major problem with X-ray images is the lack of texture, which is very important for the success of BoWs in regular images, which are rich in texture.

Table 1. Recall, precision and average precision for classification and retrieval using DoG+SIFT on low-energy, high-energy and color X-ray images

		low-energy			high-energy			color		
		Recall	Prec.	AveP	Recall	Prec.	AveP	Recall	Prec.	AveP
200	Image	0.30	0.30	0.37	0.35	0.27	0.33	0.45	0.35	0.34
	Baggage	0.40	0.22	0.42	0.40	0.15	0.37	0.70	0.28	0.44
500	Image	0.35	0.29	0.38	0.35	0.27	0.38	0.37	0.31	0.38
	Baggage	0.40	0.20	0.40	0.50	0.22	0.41	0.60	0.25	0.42
1000	Image	0.35	0.27	0.38	0.38	0.27	0.40	0.37	0.26	0.32
	Baggage	0.40	0.20	0.38	0.50	0.24	0.43	0.50	0.20	0.39

Baggage X-ray images may get very cluttered, depending on the contents of the baggage. In such cases, most of the detected keypoints belong to the background, as shown in Figure 3. We can eliminate some of the background points by a coarse background/foreground segmentation using the color information (atomic number) available in the color X-ray images. For the case of handguns, which are mostly metallic (blue in color X-ray), the foreground corresponds to blue regions. Figure 4 shows an example of such a segmentation (using Gaussian Mixture Models and morphological dilation) and the densely sampled keypoints that remain after eliminating the ones that belong to the background. Constructing the BoW representation of the images using the foreground keypoints improves the classification and retrieval performance significantly, as shown in Figure 5.

Interest point detectors detect different set of keypoints on the same image (e.g., some detect corners, some detect blobs), as shown in Figure 3. Combining the information from multiple detectors is expected to help in classification and retrieval. Indeed, as shown in Figure 6 and Table 2, using multiple point detectors

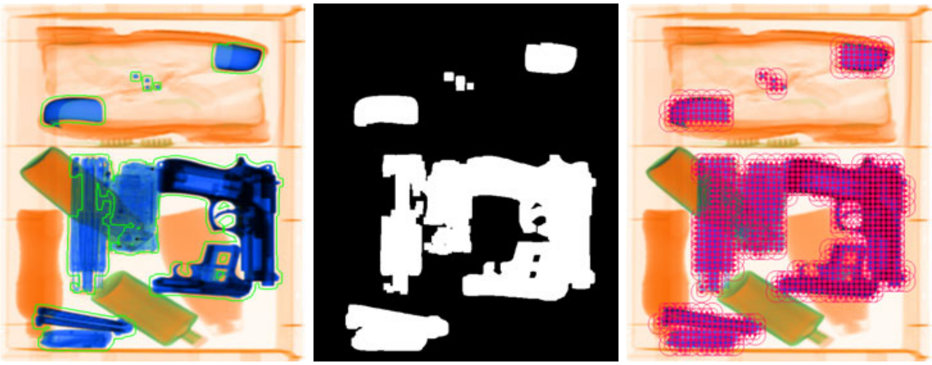


Fig. 4. Color segmentation and filtering out the background keypoints. Left: original image, middle: segmentation result for metallic materials, right: densely sampled points remaining after filtering out the points that belong to the background.

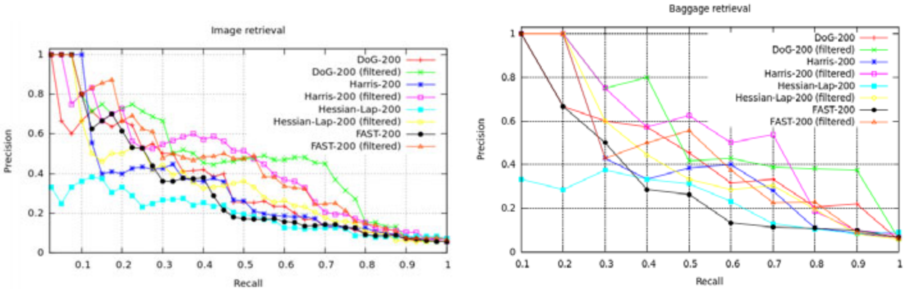


Fig. 5. Image- and baggage-based retrieval performance, with and without color filtering the keypoints. Descriptor: SIFT

Table 2. Recall, precision and average precision for classification and retrieval with multiple images per view and multiple detectors (color filtered)

		DoG (low+high+color) (DoG, union)			DoG+Harris (color, union)			DoG+Harris (color, concat.)		
		Recall	Prec.	AveP	Recall	Prec.	AveP	Recall	Prec.	AveP
200	Image	0.58	0.34	0.50	0.63	0.41	0.60	0.63	0.45	0.54
	Baggage	0.70	0.23	0.54	0.70	0.26	0.65	0.70	0.29	0.57

in computing the BoW or concatenating the BoW representation from 2 different detectors (Harris + DoG) improves the performance. Moreover, using the union of keypoints from low energy, high energy and color images also increases the performance. This suggests the use of Multiple Kernel Learning as in [2] to compute the best combination of different detectors and descriptors.

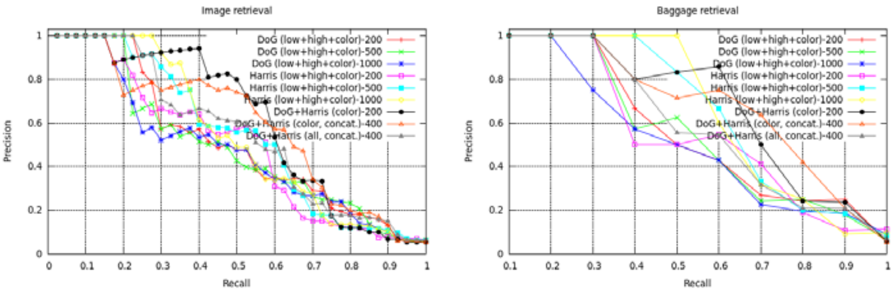


Fig. 6. Image- and baggage-based retrieval performance, with union of point detectors and concatenation of BoWs (color filtered keypoints + SIFT)

5 Conclusion

We investigated the applicability of bag of visual words method on a challenging data type, baggage X-ray images. Judging by the experimental results: (1) X-ray image data is more challenging compared to visible spectrum image data, (2) although the straightforward application of the method does not perform well, the performance can be improved by utilizing the characteristics of X-ray images (multiple views, two energy levels, color representing the material type).

Our goal is to reduce the workload of the screeners, rather than replacing them. In a realistic setting, the automatic system can process the images recorded by multiple X-ray machines and return only a small subset of the baggages as suspicious. Thus, the screeners need to examine only a small subset of the images instead of all. Hence, we also need to measure the performance of the system in terms of work reduction at specific accuracy levels.

We are currently building a large X-ray image dataset to perform a detailed benchmark for state-of-the-art classification, recognition and detection algorithms based on local features. We are planning to make the data set publicly available to stimulate research on baggage X-ray image analysis.

Acknowledgements. This work is part of the SICURA project supported by the *Bundesministerium für Bildung und Forschung* of Germany with ID FKZ 13N11125. The X-ray data is provided by Smiths-Heimann[1], a manufacturer of X-ray machines and one of the partners in the SICURA project. We thank Wan-Lei Zhao for useful discussions and Pia Dreiseitel and Won-Min Byeon for their help in acquiring the X-ray images.

References

1. OpenCV 2.2: Open source computer vision library (2011), `http://opencv.willowgarage.com`
2. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: ICCV (2009)

[1] http://www.smithsdetection.com

3. Bay, H., Ess, A., Tuytelaars, T., Gool, L.: SURF: Speeded Up Robust Features. CVIU 110(3), 346–359 (2008)
4. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: BRIEF: Binary Robust Independent Elementary Features. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 778–792. Springer, Heidelberg (2010)
5. Chang, C., Lin, C.: LIBSVM: A library for support vector machines (2011), http://www.csie.ntu.edu.tw/~cjlin/libsvm
6. Chen, Z., Zheng, Y., Abidi, B., Page, D., Abidi, M.: A Combinational Approach to the Fusion, De-noising and Enhancement of Dual-Energy X-Ray Luggage Images. In: CVPR-Workshops, p. 2 (2005)
7. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual Categorization with Bags of Keypoints. In: ECCV (2004)
8. Ding, J., Li, Y., Xu, X., Wang, L.: X-ray Image Segmentation by Attribute Relational Graph Matching. In: International Conference on Signal Processing, vol. 2 (2007)
9. He, X., Han, P., Lu, X., Wu, R.: A New Enhancement Technique of X-Ray Carry-on Luggage Images Based on DWT and Fuzzy Theory. In: International Conference on Computer Science and Information Technology, pp. 855–858 (2008)
10. Heitz, G., Chechik, G.: Object separation in x-ray image sets. In: CVPR, pp. 2093–2100 (2010)
11. Lowe, D.: Distinctive Image Features from Scale Invariant Keypoints. IJCV 60(2), 91–110 (2004)
12. Nowak, E., Jurie, F., Triggs, B.: Sampling Strategies for Bag-of-Features Image Classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 490–503. Springer, Heidelberg (2006)
13. Rebuffel, V., Dinten, J.: Dual-Energy X-Ray Imaging: Benefits and Limits. Insight-Non-Destructive Testing and Condition Monitoring 49(10), 589–594 (2007)
14. Singh, M., Singh, S.: Image Enhancement Optimization for Hand-Luggage Screening at Airports. In: Pattern Recognition and Image Analysis, pp. 1–10 (2005)
15. Sivic, J., Russell, B., Efros, A., Zisserman, A., Freeman, W.: Discovering Objects and Their Locations in Images. In: ICCV (2005)
16. Wang, L., Li, Y., Ding, J., Li, K.: Structural X-ray Image Segmentation for Threat Detection by Attribute Relational Graph Matching. In: International Conference on Neural Networks and Brain, vol. 2, pp. 1206–1211 (2006)

Image Re-ranking and Rank Aggregation Based on Similarity of Ranked Lists

Daniel Carlos Guimarães Pedronette and Ricardo da S. Torres

RECOD Lab - Institute of Computing - University of Campinas,
CEP 13083-970, Campinas/SP - Brazil

Abstract. The objective of Content-based Image Retrieval (CBIR) systems is to return a ranked list containing the most similar images in a collection given a query image. The effectiveness of these systems is very dependent on the accuracy of the distance function adopted. In this paper, we present a novel approach for redefining distances and later re-ranking images aiming to improve the effectiveness of CBIR systems. In our approach, distance among images are redefined based on the similarity of their ranked lists. Conducted experiments involving shape, color, and texture descriptors demonstrate the effectiveness of our method.

1 Introduction

Traditional image retrieval approaches based on keywords and textual metadata face serious challenges. Describing the image content with textual descriptions is intrinsically very difficult [6]. One of the most common approaches to overcome these limitations on supporting image searches relies on the use of Content-Based Image Retrieval (CBIR) systems. Basically, given a query image, a CBIR system aims at retrieving the most similar images in a collection by taking into account image visual properties (such as, shape, color, and texture). Collection images are *ranked* in decreasing order of similarity, according to a given image descriptor.

Commonly, CBIR systems compute similarity considering only pair of images. On the other hand, the user perception usually considers the query specification and responses in a given *context*. For interactive applications, the use of context can play an important role [1]. In a CBIR scenario, relationships among images and information encoded in ranked lists can be used for extracting contextual information.

Recently, efforts were put on post-processing the similarity scores aiming to improve the effectiveness of information retrieval tasks [13]. Methods have been proposed for analyzing the relations among all documents in a given collection [8,25,24,14]. *Contextual information* have also been considered for improving the effectiveness of image retrieval approaches [15,18,13]. The objective of these methods is somehow mimic the human behavior on judging the similarity among objects by considering specific *contexts*. More specifically, the notion of *context* can refer to updating image similarity measures by taking into account information encoded in the ranked lists defined by a CBIR system [18].

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 369–376, 2011.

In this paper, we present the *RL-Sim Re-Ranking Algorithm*, a new post-processing method that considers *ranked lists similarities* for taking into account contextual information. We propose a novel approach for computing new distances among images based on the similarity of their ranked lists. Each ranked list is modeled as *sets* and set operations are used for computing the similarity between two ranked lists. We believe that the modeling of ranked lists as sets, in a general way, represents an advantage of our strategy. The algorithm can be used with different distances or similarity scores. Thus, the re-ranking method can be used for different CBIR tasks and easily adapted for other information retrieval tasks (e.g., text or multimodal search).

We evaluated the proposed method with shape, color, and texture descriptors. Experimental results demonstrate that the proposed method can be used in several CBIR tasks and yields better results in terms of effectiveness performance than various post-processing algorithms recently proposed in the literature.

2 The RL-Sim Re-ranking Algorithm

The main motivation of our re-ranking algorithm relies on the conjecture that *contextual information encoded in the similarity between ranked lists can provide resources for improving the effectiveness of CBIR descriptors*. In general, if two images are similar, their ranked lists should be similar as well [14]. It is somehow close to the the cluster hypothesis [16], which states that *"closely associated documents tend to be relevant to the same requests"*. In the following, we present a definition of our re-ranking algorithm.

Let $\mathcal{C}$=$\{img_1, img_2, \ldots, img_N\}$ be an image collection and let $\mathcal{D}$ be an image descriptor that defines a distance function $\rho : \mathcal{C} \times \mathcal{C} \rightarrow \mathbb{R}$, where $\mathbb{R}$ denotes real numbers. Consider $\rho(x, y) \geq 0$ for all (x, y) and $\rho(x, y) = 0$, if $x = y$. The distance $\rho(img_i,img_j)$ among all images img_i,$img_j \in \mathcal{C}$ can be computed to obtain an $N \times N$ distance matrix A.

Given a query image img_q, we can compute a ranked list R_q in response to the query, based on the distance matrix A. The ranked list R_q=$\{img_1$, img_2, $\ldots$, $img_N\}$ can be defined as a permutation of the collection $\mathcal{C}$, such that, if img_1 is ranked at lower positions than img_2, then $\rho(img_q,img_1) \leq \rho(img_q,img_2)$. We also can take every image $img_i \in \mathcal{C}$ as a query image img_q, in order to obtain a set $\mathcal{R} = \{R_1, R_2, \ldots, R_N\}$ of ranked lists.

Let $\psi : \mathcal{R} \times \mathcal{R} \times \mathbb{N} \rightarrow \mathbb{R}$ be a similarity function that defines a similarity score between two ranked lists considering their first K images, such that $\psi(R_x, R_y, K) \geq 0$ for all $R_x, R_y \in \mathcal{R}$.

Our goal is to propose a re-ranking algorithm that takes as input an initial set of ranked lists $\mathcal{R}$ and use the function ψ for computing a more effective distance matrix $\hat{A}$ and therefore a more effective set of ranked lists $\hat{\mathcal{R}}$. An iterative approach is proposed: the new (and more effective) set $\mathcal{R}_{t+1}$, where t indicates the current iteration, is used for the next execution of our re-ranking algorithm and so on. These steps are repeated along several iterations aiming to improve the effectiness incrementally. After a number T of iterations a re-ranking is performed based on the final distance matrix $\hat{A}$. Based on matrix $\hat{A}$, a final set

Algorithm 1. RL-Sim Re-Ranking Algorithm

Require: Original set of ranked lists $\mathcal{R}$, distance matrix A, and parameters K_s, T, λ
Ensure: Processed set of ranked lists $\hat{\mathcal{R}}$

```
 1: $t \leftarrow 0$
 2: $\mathcal{R}_t \leftarrow \mathcal{R}$
 3: $A_t \leftarrow A$
 4: $K \leftarrow K_s$
 5: while $t < T$ do
 6:   for all $R_i \in \mathcal{R}_t$ do
 7:     $c \leftarrow 0$
 8:     for all $img_j \in R_i$ do
 9:       if $c \leq \lambda$ then
10:         $A_{t+1}[i,j] \leftarrow 1/(1 + \psi(R_i, R_j, K))$
11:       else
12:         $A_{t+1}[i,j] \leftarrow 1 + A_t[i,j]$
13:       end if
14:       $c \leftarrow c + 1$
15:     end for
16:   end for
17:   $\mathcal{R}_{t+1} \leftarrow perfomReRanking(A_{t+1})$
18:   $t \leftarrow t + 1$
19:   $K \leftarrow K + 1$
20: end while
21: $\hat{\mathcal{R}} \leftarrow \mathcal{R}_T$
```

of ranked lists $\hat{\mathcal{R}}$ can be computed. Algorithm 1 outlines the proposed *RL-Sim Re-Ranking Algorithm.*

Observe that the distances are redefined considering the function ψ for the first λ positions of the each ranked list, such that $\lambda \in \mathbb{N}$ and $0 \leq \lambda \leq N$. For images in the remaining positions of the ranked lists, the new distance is redefined (Line 12) based on the current distances. In these cases, the function ψ does not need to be computed, considering that relevant images should be at the begining of the ranked lists. In this way, the computional efforts decreases, becoming the algorithm not dependent on the collection size N.

Also note in Line 19, that at each iteration t, we increment the number of K neighbors considered in the computation of function ψ. The motivation behind this increment relies on the fact that the effectiveness of ranked lists increase along iterations. In this way, non-relevant images are moved out from the first positions of the ranked lists and K can be increased for considering more images.

The main step of the algorithm consists in the computation of function ψ, detailed in next subsection.

2.1 Measuring Similarity between Ranked Lists

The objective of function ψ is to compute a more effective distance between two images, considering the contextual information encoded in the first K positions of their ranked lists. Let KNN be a function that extracts a subset with the first positions of ranked list R_i, such that $\mid KNN(R_i, k) \mid = k$. The function ψ computes the intersection between the subsets of two ranked lists considering different values of k, such that $k \leq K$. Equation 2 formaly defines the function ψ.

$$\psi(R_x, R_y, K) = \frac{\sum_{k=1}^{K} \mid KNN(R_x, k) \cap KNN(R_y, k) \mid}{K} \tag{1}$$

Note that if two ranked lists present the same images at the first positions, the size of the intersection set is greater, and the value of ψ is greater as well. Figure 1 illustrates the computation of ψ considering multiscale values of K.

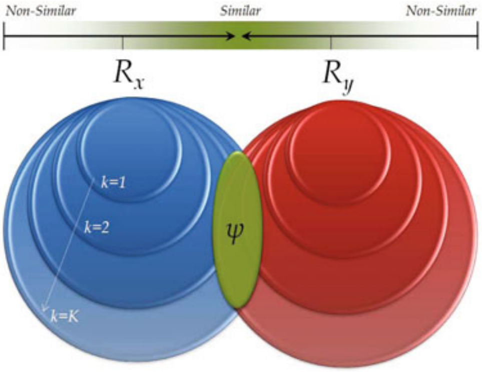


Fig. 1. Computation of function ψ: intersection of ranked lists with different sizes

3 The Rank Aggregation Algorithm

Let $\mathcal{C}$ be an image collection and let $\mathcal{D} = \{D_1, D_2, \ldots, D_m\}$ be a set of CBIR descriptors. We can use the set of descriptors $\mathcal{D}$ for computing a set of distances matrices $\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$. Our approach for combining descriptors works as follows: first, we combine the set $\mathcal{A}$ in a unique matrix A_c. For the matrices combination we use a multiplicative approach. Each position (i, j) of the matrix is computed as follows:

$$A_c[i,j] = A_1[i,j] \times A_2[i,j] \times \ldots A_m[i,j] \tag{2}$$

Once we have a matrix A_c, we can compute a set of ranked lists $\mathcal{R}_c$ based on this matrix. Then, we can submit the matrix A_c and the set $\mathcal{R}_c$ for our original re-ranking algorithm.

4 Experimental Evaluation

4.1 Impact of Parameters

The execution of Algorithm 1 considers three parameters: *(i)* K_s - number of neighbors considered when algorithm starts; *(ii)* λ - number of images of each ranked list that are considered for redefining distances; and *(ii)* T - number of iterations in which the algorithm is executed.

To evaluate the influence of different parameter settings on the retrieval scores and for determining the best parameters values we conducted a set of experiments. We use MPEG-7 database with the so-called bullseye score, which counts all matching objects within the 40 most similar candidates. Since each class consists of 20 objects, the retrieved score is normalized with the highest possible number of hits. For distance computation, we used the ASC [11] shape descriptor.

Retrieval scores are computed ranging parameters K_s in the interval [1,20] and T in the interval [1,7]. Figure 2 illustrates the results of precision scores for different values of parameters K_s and T. We observed that best retrieval scores increased along iterations and parameters converged for values $K_s = 15$ and $T = 3$: 94.69%. We used these values in all experiments.

We also analyzed the impact of parameter λ on precision. As discussed before, the objective of λ consists in decreasing computation efforts needed for the algoritm. In this way, we ranged λ in the interval [0,N] (considering the MPEG-7 collection). Results are illustrated in Figure 3. Note that the precision scores achieve the stability near to $\lambda = 700$ (value used in our experiments).

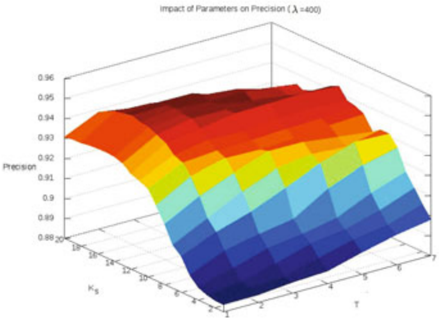


Fig. 2. Impact of parameters: K_s and T

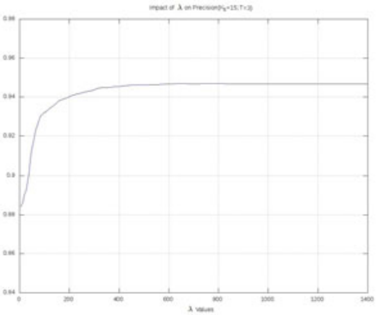


Fig. 3. Impact of parameters: λ

4.2 Experiments Considering CBIR Tasks

In this section, we present a set of conducted experiments for demonstrating the applicability and effectiveness of our method. We analyzed our method with in the task of re-ranking images considering shape, color, and texture descriptors. We also compared our method to state-of-the-art post-processing methods.

Table 1 presents results (bullseye score - Recall@40) for shape descriptors on MPEG-7 database. We can observe a significative gains from +7.13% to +20.82%.

In addition to shape descriptors, we conducted experiments with color and texture descriptors. For texture descriptor, we used the Brodatz [5] dataset, a popular dataset for texture descriptors evaluation. For color descriptor, we used a soccer data set proposed in [23] and composed by images from 7 soccer teams, containing 40 images per class. Table 2 presents results for 10 image descriptors in 3 different datasets. The measure adopted is *Mean Average Precision (MAP)*. We can observe that the proposed re-ranking method presented gains of up to +15% in MAP scores.

Table 1. Contextual Re-Ranking for shape descriptors on MPEG-7 *(Recall@40)*

Shape Descriptor	**Score [%]**	**RL-Sim Re-Ranking**	**Gain**
SS [17]	43.99%	53.15%	+20.82%
BAS [2]	75.20%	82.94%	+10.29%
IDSC+DP [10]	85.40%	92.18%	+7.94%
CFD [14]	84.43%	94.13%	+11.49%
ASC [11]	88.39%	94.69%	+7.13%

Table 2. MAP scores regarding the used of the RL-Sim Re-Ranking in CBIR tasks

Descriptor	**Type**	**Dataset**	**Score** (MAP)	**RL-Sim Re-Ranking** (MAP)	**Gain**
SS [17]	Shape	MPEG-7	37.67%	43.06%	+14.31%
BAS [2]	Shape	MPEG-7	71.52%	74.57%	+4.25%
IDSC [10]	Shape	MPEG-7	81.70%	86.75%	+6.18%
ASC [11]	Shape	MPEG-7	85.28%	88.81%	+4.14%
CFD [14]	Shape	MPEG-7	80.71%	88.97%	+10.23%
GCH [20]	Color	Soccer Dataset	32.24%	33.66%	+4.40%
ACC [7]	Color	Soccer Dataset	37.23%	43.54%	+16.95%
BIC [19]	Color	Soccer Dataset	39.26%	43.45%	+10.67%
LBP [12]	Texture	Brodatz	48.40%	47.77%	-1.30%
CCOM [9]	Texture	Brodatz	57.57%	62.01%	+7.72%
LAS [21]	Texture	Brodatz	75.15%	77.81%	+3.54%

We evaluated the use of our re-ranking method to combine different CBIR descriptors. We selected two descriptors for each visual property: descriptors with best effectiveness results are selected. Table 3 presents results of MAP score of these descriptors. We can observe significatve gains when compared with the results obtained for descriptor in isolation.

Table 3. MAP scores regarding the use of RL-Sim Algorithm for Rank Aggregation

Descriptor	**Type**	**Dataset**	**Score** (MAP)
CFD [14] + IDSC [10]	Shape	MPEG-7	98.34%
CFD [14] + ASC [11]	Shape	MPEG-7	98.75%
BIC [19] + ACC [7]	Color	Soccer	44.49%
LAS [21] + CCOM [9]	Texture	Brodatz	80.26%

Finally, we also evaluated our method in comparison to other state-of-the-art post-processing methods. We use MPEG-7 database with the called bullseye score. Table 4 presents results of our contextual re-ranking method and four post-processing methods. Note that the results of our **RL-Sim Re-Ranking** method in rank aggregation tasks presented the best effectiveness performace when compared to other methods.

Table 4. Post-processing methods comparison on MPEG-7 database *(Recall@40)*

Algorithm	**Shape Descriptor**	**Score**	**Gain**
Data Driven Generative Models (DDGM) [22]	-	80.03%	-
Contour Features Descritpor (CFD) [14]	-	84.43%	-
Inner Distance Shape Context (IDSC) [10]	-	85.40%	-
Shape Context (SC) [4]	-	86.80%	-
Aspect Shape Context (ASC) [11]	-	88.39%	-
Graph Transduction (LP) [24]	IDSC [10]	91.00%	+6.56%
Distance Optimization [14]	CFD [14]	92.56%	+9.63%
Locally Constrained Diffusion Process [25]	IDSC [10]	93.32%	+9.27%
Mutual kNN Graph [8]	IDSC [10]	93.40%	+9.37%
RL-Sim Re-Ranking	**CFD [14]**	**94.13%**	**+11.49%**
Contextual Re-Ranking [13]	CFD [14]	94.55%	+11.99%
RL-Sim Re-Ranking	**ASC [11]**	**94.69%**	**+7.13%**
Locally Constrained Diffusion Process [25]	ASC [11]	95.96%	+8.56%
Co-Transduction [3]	IDSC [10]+DDGM [22]	97.31%	-
Co-Transduction [3]	SC [4]+DDGM [22]	97.45%	-
Co-Transduction [3]	SC [4]+IDSC [10]	97.72%	-
RL-Sim Re-Ranking	**CFD [14]+IDSC [10]**	**99.31%**	-
RL-Sim Re-Ranking	**CFD [14]+ASC [11]**	**99.44%**	-

5 Conclusions

In this work, we have presented a new re-ranking method that exploits contextual information for improving CBIR tasks. The main idea consists in analyzing similarity between ranked lists for redefing distance among images. We conducted a large set of experiments and experimental results demonstrated the applicability of our method to several image retrieval tasks based on shape, color, and texture descriptors. Future work focuses on considering multimodal searches involving visual and textual descriptions associated with images.

Acknowledgment. Authors thank FAEPEX, CAPES, FAPESP, and CNPq for financial support. Authors also thank DGA/UNICAMP for its support in this work.

References

1. Abowd, G.D., Dey, A.K., Brown, P.J., Davies, N., Smith, M., Steggles, P.: Towards a better understanding of context and context-awareness. In: Gellersen, H.-W. (ed.) HUC 1999. LNCS, vol. 1707, pp. 304–307. Springer, Heidelberg (1999)
2. Arica, N., Vural, F.T.Y.: Bas: a perceptual shape descriptor based on the beam angle statistics. Pattern Recogn. Lett. 24(9-10), 1627–1639 (2003)
3. Bai, X., Wang, B., Wang, X., Liu, W., Tu, Z.: Co-transduction for shape retrieval. In: ECCV, vol. 3, pp. 328–341 (2010)
4. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. PAMI 24(4), 509–522 (2002)
5. Brodatz, P.: Textures: A Photographic Album for Artists and Designers. Dover, New York (1966)
6. Faria, F.F., Veloso, A., Almeida, H.M., Valle, E., da, S., Torres, R., Gonçalves, M.A., Meira Jr., W.: Learning to rank for content-based image retrieval. In: MIR 2010, pp. 285–294 (2010)

7. Huang, J., Kumar, S.R., Mitra, M., Zhu, W.J., Zabih, R.: Image indexing using color correlograms. In: CVPR 1997, p. 762 (1997)
8. Kontschieder, P., Donoser, M., Bischof, H.: Beyond pairwise shape similarity analysis. In: Zha, H., Taniguchi, R.-i., Maybank, S. (eds.) ACCV 2009. LNCS, vol. 5996, pp. 655–666. Springer, Heidelberg (2010)
9. Kovalev, V., Volmer, S.: Color co-occurence descriptors for querying-by-example. In: MMM 1998, p. 32 (1998)
10. Ling, H., Jacobs, D.W.: Shape classification using the inner-distance. PAMI 29(2), 286–299 (2007)
11. Ling, H., Yang, X., Latecki, L.J.: Balancing deformability and discriminability for shape matching. In: ECCV, vol. 3, pp. 411–424 (2010)
12. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. PAMI 24(7), 971–987 (2002)
13. Pedronette, D.C.G., da S. Torres, R.: Exploiting contextual information for image re-ranking. In: Bloch, I., Cesar Jr., R.M. (eds.) CIARP 2010. LNCS, vol. 6419, pp. 541–548. Springer, Heidelberg (2010)
14. Pedronette, D.C.G., da, S., Torres, R.: Shape retrieval using contour features and distance optmization. In: VISAPP, vol. 1, p. 197–202 (2010)
15. Perronnin, F., Liu, Y., Renders, J.M.: A family of contextual measures of similarity between distributions with application to image retrieval. In: CVPR, pp. 2358–2365 (2009)
16. van Rijsbergen, C.: Information Retrieval (1979)
17. da, S., Torres, R., Falcão, A.X.: Contour Salience Descriptors for Effective Image Retrieval and Analysis. Image and Vision Computing 25(1), 3–13 (2007)
18. Schwander, O., Nielsen, F.: Reranking with contextual dissimilarity measures from representational bregmanl k-means. In: VISAPP, vol. 1, pp. 118–122 (2010)
19. Stehling, R.O., Nascimento, M.A., Falcão, A.X.: A compact and efficient image retrieval approach based on border/interior pixel classification. In: CIKM 2002, pp. 102–109 (2002)
20. Swain, M.J., Ballard, D.H.: Color indexing. IJCV 7(1), 11–32 (1991)
21. Tao, B., Dickinson, B.W.: Texture recognition and image retrieval using gradient indexing. JVCIR 11(3), 327–342 (2000)
22. Tu, Z., Yuille, A.L.: Shape matching and recognition - using generative models and informative features. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3023, pp. 195–209. Springer, Heidelberg (2004)
23. van de Weijer, J., Schmid, C.: Coloring local feature extraction. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 334–348. Springer, Heidelberg (2006)
24. Yang, X., Bai, X., Latecki, L.J., Tu, Z.: Improving shape retrieval by learning graph transduction. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 788–801. Springer, Heidelberg (2008)
25. Yang, X., Koknar-Tezel, S., Latecki, L.J.: Locally constrained diffusion process on locally densified distance spaces with applications to shape retrieval. In: CVPR, pp. 357–364 (2009)

A Cartography of Spatial Relationships in a Symbolic Image Database

Nguyen Vu Hoàng[1,2], Valérie Gouet-Brunet[2], and Marta Rukoz[1,3]

[1] LAMSADE - U. Paris-Dauphine
[2] CEDRIC - CNAM
[3] U. Paris Ouest Nanterre La Défense

Abstract. This work addresses the problem of the representation of spatial relationships between symbolic objects in images. We have studied the distribution of several categories of relationships in LabelMe[1], a public database of images where objects are annotated manually and online by users. Our objective is to build a cartography of the spatial relationships that can be encountered in a representative database of images of heterogeneous content, with the main aim of exploiting it in future applications of Content-Based Image Indexing (CBIR), such as object recognition or retrieval. In this paper, we present the framework of the experiments made and give an overview of the main results obtained, as an introduction to the website[2] dedicated to this work, whose ambition is to make available all these statistics to the CBIR community.

Keywords: Symbolic image, Spatial relationships, Cartography.

1 Introduction

We are interested in the representation of spatial relationships between symbolic objects in images. In CBIR, embedding such information into image content description provides a better representation of the content as well as new scenarios of interrogation. Literature on spatial relationships is very rich – several hundreds of papers exist on this topic – and a lot of approaches were proposed. Most of them describe different aspects of spatial relationships, e.g. directional [3] or topological [1] relationships, and have been evaluated on small synthetic or specific image datasets, e.g. medical or satellite imagery. In this work, we propose to build a cartography of the spatial relationships that can be encountered in a database of images of heterogeneous natural contents, such as audiovisual, web or family visual contents. We have chosen a public annotated database, from the platform LabelMe[1], which is described in section 2. This cartography collects statistical informations on the trends of spatial relationships involving symbolic objects effectively encountered in this database, with the aim of exploiting them in future CBIR applications, for improving tasks such as object recognition or

[1] LabelMe: http://labelme.csail.mit.edu

[2] Our website: http://www.lamsade.dauphine.fr/~hoang/www/cartography

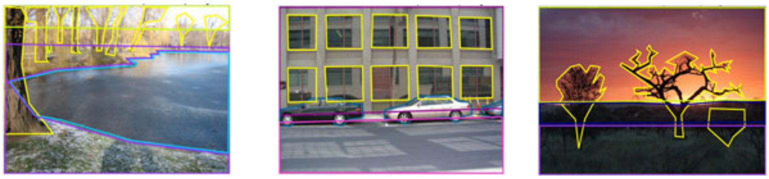
A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 377–385, 2011.

retrieval. Here, we focus on the analysis of unary and binary relationships, respectively described in sections 3 and 4. Because of space limitation, we present a synthesis of this analysis, which is fully described in research report [2], and made available to the CBIR community on our website[2].

2 A Public Annotated Image Database: LabelMe

Studied database: LabelMe [7] is a platform containing image databases and an online annotation tool that allows users to indicate freely, by constructing a polygon and a label, the many objects depicted in a image as they wish. Thus, each object, called entity in this work, is presented by a polygon and a label. In our work, each label is considered as the name of an entity category, so all entities possessing the same label belong to a same category. We used one of the test databases of this platform which contains 1133 annotated images in daily contexts (see examples in Fig.1). The content of these images is very heterogeneous, it contains many categories and many images, and it is not specific to a particular domain. Therefore, studying this database can provide a general view about categories and their relationships, and the results should not be influenced noticeably by changing the database. In order to guarantee the quality of the database we verified carefully each annotated image for consistency. Firstly, we manually consolidated synonymous labels by correcting orthographic mistakes and merging labels having the same meaning. Secondly, we selected 86 different categories taking into account only those categories having at least 15 entities. This decision was taken to ensure an independence of statistical results even whether the image database is changed. Lastly, we added missing annotations to entities of the considered categories, except for too small size entities or entities belonging to a category having a high frequency of already annotated entities in the image. In this way, the statistical results should not be biased by these missing annotations. In the following, we call DB this database.

Statistic on categories: Before studying different relationships between categories, we take a look at statistics concerning each category, for example, its highest and lowest numbers of entities in an image, the total number of its entities in the database, the number of images where at less one of its entities appears, etc. The overview of this statistical study is presented in Table 1.



Sky, tree, person, lake, ground

Road,car, building, window

Sky,tree, mountain, ground

Fig. 1. Images of DB and associated annotations

Table 1. Overview of statistical measures on DB

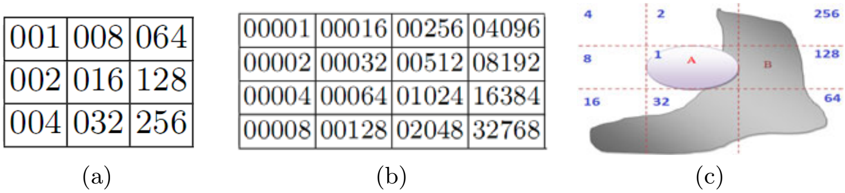
Nb of img/DB	Nb of entities/DB	Average of entities/cat.(STDEV)	Average of entities/img (STDEV)	Max. nb of entities/img	Min. nb of entities/img
1133	38075	442.7 (1485.6)	33.6 (32.3)	264	1

The average entities number of each category in an image could be used to have a quick view about the possibility of having more than one of its entities in an image. For example, the WINDOW category has a high average, around 19 entities per image. Then, if we find a WINDOW in an image, we can expect to find another WINDOW in the same image. For a more detailed study, we have computed the intra-class correlation of categories, based on the classic correlation function between two categories. Slightly differently to classic correlation representing impact of one category on another, the intra-class correlation is never negative. Returning to the previous example, we obtained 0.776 for the intra-class correlation of WINDOWS, this is also the highest score among intra-class correlations obtained. This score is high enough to conclude that we can find mostly at least two WINDOWS in an image where a WINDOW entity has already been detected. The lowest score in this study is 0, related to LAKE category. Therefore, no image in DB contains more than one LAKE. In fact, it is not usual to have two or more instances of LAKE in the same image. Summarizing, 21 categories have intra-class correlation higher than 0.3 while only 8 categories have a score higher than 0.5, for example CAR, WINDOW, BUILDING.

This study can provide useful information in the category detection process, if we want, for example, to detect all entities of a category C_i present in an image I. Knowing that C_i has, in general, one entity per image (based on a threshold on correlation, for example), as soon as the first entity of C_i is detected, we could finish the detection process, thus reducing significantly the execution time of the detection. The statistics for all categories are available on our website[2].

3 Unary Relationships

Representation: We call unary relationship, the relationship between an entity and its localization in an image, where localization is defined as a region or area of the image, represented in this work by a code. More formally, let $A = \{A_i\}$, $I = \{I_j\}$, and $C = \{C_k\}$ be the set of areas, the set of images, and the set of categories, respectively. The unary relationship is an application R from $C \times I$ to A. $R(C_k, I_j) \in A$ allows knowing where C_k is located in I_j. Areas of an image can be represented in different ways like quad-tree or quin-tree, see [6,8]. Since we do not have any knowledge a priori of the location of the categories in the images, we propose to split images in a fixed number of regular areas (i.e. equal size areas). First, we divide each image in a fixed sized grid. Each cell of this grid, called atomic area, is represented by a code. Fig.2(a) and 2(b) depict a splitting in 9 and 16 atomic areas and their codes, respectively. We then combine these codes to present more complex areas, by example for 9-area splitting, the code 009 represents area ▦ grouping together atomic areas 001(▦) and 008(▦).



001	008	064
002	016	128
004	032	256

(a)

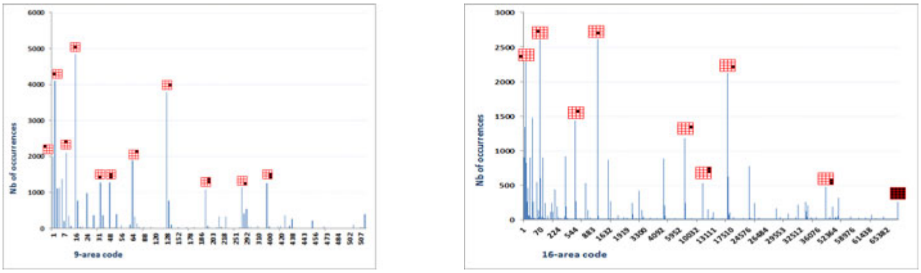
00001	00016	00256	04096
00002	00032	00512	08192
00004	00064	01024	16384
00008	00128	02048	32768

(b)

(c)

Fig. 2. Codes in (a) 9-area splitting, (b) 16-area splitting, (c) 9DSpa areas

Results analysis: The combination of nine codes in 9-area splitting (Fig.2(a)) gives 511 possible complex area codes. However, some of them cannot be used, for example, code 017 (▦) or code 161 (▦) because their atomic areas are not connected by an edge (i.e. they are disjoint). In consequence, based on this idea, there are only 218 theoretically authorized codes. Concretely, in DB, we did not find any entity in areas represented by impossible codes. Moreover, there are only 138 useful theoretically authorized codes. For example, DB does not contain any entity in areas with codes 047(▦) or 125 (▦). In the same way, the combination of 16 codes in Fig.2(b) gives 65535 different codes. In theory, we can reach 11506 possible complex areas (based on connected areas), but in DB, only 649 codes are present. A quick report about codes present in DB is presented in Fig.3.



(a)

(b)

Fig. 3. Distribution of codes in terms of nb. of occurrences: (a)9-area splitting (b)16-area splitting.

Interpretation: From Fig.3(b) we can observe that large or complex regions have a small number of occurrences. It means that categories are mostly represented by a simple or small area. On the other hand, the trend of the categories' presence first corresponds to the two middle lines, then to the combination of the two top lines, and finally to the combination of the two bottom lines. These results are consistent with those of 9-area splitting (see Fig.3(a)). Similarly, we can observe that the trend of the categories' presence, on the left is higher than on the right. These conclusions confirm the well known rules concerning photography and ergonomics (human-computer interaction):

- In photography, there is the rule of thirds [3], one of the first rules of composition taught to most photography students. It is recommended to present interesting objects at the intersections or along the lines presented in this rule (see Fig.3(b)).
- According to [4] concerning ergonomic studies on human-computer interaction, the center of computer screen is the most attracting. Next, the human attention view is attracted by the top and the left of screen more than by the bottom and the right consecutively, leading to slightly more annotated entities in these areas (see Fig.3(a)).

We have studied the distribution of categories across areas of the image, according to 9-area and 16-area splitting. Basically, the results obtained can be encapsulated in a knowledge-based system where they will be interpreted as a probability of presence of a given category in a given area. For example with 9-area splitting, CHIMNEY and SKY appear more frequently on the top of the image (of probabilities 0.72 and 0.81 respectively). In a object detection task for example, these measures can help in determining priority searching areas, and then in reducing the searching space of the objects. They are available for all categories on our website[2].

Spatial reasoning: We have also a question: *"Could a category be frequently and entirely present in a given area?"*. This question could help us to find an efficient method for detecting a category in an image. This idea drives us to examine the distribution of occurrences of each category C_j according to each theoretically possible area in the image, by the way of a normalized histogram: $H_{split}(C_j)$ with *split* $\in$ {9-area, 16-area}. When a category is integrally in an area A_i of *split*, it can probably appear in a smaller theoretically authorized area A_k included in A_i. Let FC be a function allowing to create theoretically possible areas from A_i. $\{A_k\} = FC_{split}(A_i)$. Let $SC_{split}(A_i)$ be the set of codes of every theoretically authorized areas A_k in A_i: $SC_{split}(A_i) = \{cod(A_k)|A_k \in FC_{split}(\{A_{split}\})\}$, where $cod(A_k)$ is the code representing area A_k. A category C_j, whose instances appear in A_i, has a specific histogram where the number of occurrences of a code c, $c \in SC_{split}(A_i)$, is not null. Then, to do spatial reasoning on such histograms, we propose a function FH such as $FH(H_{split}(C_j), A_i) = G_{split} \odot H_{split}(C_j)$, where $\odot$ is the dot product and G a 1D template mask of size the number of theoretical codes c according to the splitting method:

$$G_{split}(c) = \begin{cases} 0 \text{ if } c \in SC_{split}(A_i) \\ 1 \text{ otherwise} \end{cases} \tag{1}$$

FH has values varying in [0..1] ; $FH = 0$ means that all not null frequencies correspond to codes $SC_{split}(A_i)$, and then that category C_j is always entirely in area A_i. If $FH = 1$, we can say that C_j is never entirely in A_i. The more FH is high, the less C_j appears entirely in A_i. From FH, we can deduce the probability p_a of presence of C_j in A_i as $p_a(A_i) = 1 - FH$. More generally, if we examine the presence of C_j in n disjoint areas A_i, the probability becomes

[3] http://www.digital-photography-school.com/rule-of-thirds

$p_a(\{A_i\}_n) = \sum_{i=1}^{n}(1 - FH(H_{split}(C_j), A_i))$. For category PERSON for example, the values FH for the three A_i areas ▦ ▦ ▦ in 9-area splitting are respectively 0.704, 0.644 and 0.721, that gives $p_a(\{A_i\}_3) = 0.931$. This result means that the probability of category PERSON to be entirely in one column is high, and that its presence in two columns at least is very small. Consequently, we can say that in DB, entities PERSON are present vertically most of the time, and that they appear rarely at scales larger than one column. These statistics can help designing a person detection task for future applications. Similar spatial reasoning can be done with other categories and other areas.

4 Binary Relationships

A binary relationship links two entities of distinct categories together in an image. It can be a co-occurrence or a spatial relationship. From the 86 categories of the database used, there are 3655 possible binary relationships between categories. Among them, we observed first that 879 couples of categories never occur together. Before studying spatial binary relationships, we examine co-occurrence relationships.

Co-Occurrence Relationships

To begin, we give an example. In DB, WINDOW appears in 677 images and CAR in 519 images, while the couple appears together in 480 images. Then, we can conclude that their co-occurrence relationship is quite remarkable: for instance, 92% of the images containing a CAR also contain a WINDOW. This rate corresponds to a conditional probability, denoted $P(\text{WINDOW}|\text{CAR})$. Additionally, we can compute their correlation to learn more about the co-occurrence of such couples. Hence, these measures can help understanding better which category's presence conducts to the presence or absence of another category. Because of article's length limitation, we present only some relevant statistics in Table 2.

Interpretation: the correlation score can resume in one value the presence or absence together of two categories and especially the strength of this knowledge.

Table 2. Couples of categories having either the highest or lowest number of occurrences, of conditional probability or of correlation

Couple (A-B)	Nb of occur.	$P(A)$	$P(B)$	$P(A \cap B)$	$P(A\|B)$	$P(B\|A)$	Corr.
WINDOW-CAR	57925	0.598	0.458	0.424	0.925	0.709	0.609
BUILDING-SIDEWALK	3051	0.688	0.542	0.535	0.987	0.777	0.696
WINDOW-BUILDING	38173	0.598	0.688	0.591	0.859	0.990	0.788
WINDOW-LAKE	0	0.598	0.014	0.000	0.000	0.000	-0.149
CAR-TAIL LIGHT	1591	0.458	0.147	0.147	1.000	0.321	0.450
CHIMNEY-SKY	78	0.040	0.653	0.040	0.061	1.000	0.146
BUILDING-BIRD	15	0.688	0.046	0.004	0.077	0.005	-0.297
ARM-TORSO	2262	0.089	0.092	0.089	0.971	1.000	0.984

The highest score obtained is 0.984 for (TORSO-ARM) ; in fact, only 3 couples have a correlation higher than 0.8. The lowest score obtained is −0.297 for (BUILDING-BIRD). These results cannot conduct to the conclusion on correlation or decorrelation of most of the couples of categories. But conditional probabilities can help to go deeper in the analysis. For example P(BUILDING|SIDEWALK) is very high (see Table 2). That means that, in detecting a SIDEWALK, we can expect finding a BUILDING in the same image. Such relationship should be integrated with benefit in a knowledge-based system dedicated to artificial vision. Indeed, sidewalks are easy to detect because of their specific and universal visual appearance, while the variability of buildings makes them harder to detect. Then the prior detection of a sidewalk would contribute to facilitate the detection of a building by reducing the number of images to process. This reasoning can be generalized to other couples of categories, since in total, there are 141 conditional probabilities higher than 0.95. Note that 66 of them are equal to 1 (see examples in Table 2), making the possibility of replacing the detection step of one category by the detection step of another, if easier, to find images of that category. All these measures are available on the website[2] of this work.

Binary Spatial Relationships

In last years, there have been many approaches proposed for representing binary spatial relationships. They can be classified as topological, directional or distance-based approaches, and can be applied on symbolic objects or low level features. Here, we have focussed on relationships between the entities of DB described in terms of directional relationships with approach 9DSpa [3], of topological relationships [1] and of a combination of them with 2D projections [5]. 9DSpa describes directional relationships between a reference entity and another one based on the combination of 9 codes associated to areas orthogonally built around the MBR (Minimum Bounding Rectangle) of the reference entity ; see the illustration of Fig. 2(c) where the reference entity A has a relationship with B of code 499 and graphical representation ▦. The description of topological relationships produces 8 types of relationships. The 2D projections approach associates 7 basic operators plus 6 symmetric ones (denoted by adding symbol "*" to the basic ones) to each image axis, leading to 169 possible 2D relationships between MBR of entities. Table 3 presents an overview on the statistics obtained in terms of relationships encountered and of occurrences in DB.

Interpretation: among all the possible relationships existing theoretically, only a subset was effectively found in DB for each approach. The subset is

Table 3. Binary spatial relationships studied and related main statistics

Approach	Nb of possible relationships	Nb of effective relationships	Relationships with best occurrences (and frequency in %)
9DSpa	511	206	◰ (14%), ◱ (13%), ◲ (14%), ◳(13%)
Topological rel.	8	5	"Disjoint" (94%)
2D projections	169	36	$<$ (37%), $<*$ (37%) *(averaged on x,y axes)*

particularly small with 9DSpa and 2D projections (see Table 3). This result leads to the first conclusions that the digital codes of these relationships could be optimized and that indexing them would more benefit from data driven than space driven indexes. With topological relationships, we never found ”equal”, ”cover” and ”cover by”. ”Meet” relationship is dully represented (0.3%): its number of occurrences is small because the notion of strict adjacency between high-level objects is not common in natural contents such those of DB and because of manual annotation. Meanwhile, in literature ”meet” is a popular relationship often used with some image analysis techniques such as region segmentation that generates adjacent regions by definition, with application to specific domains, e.g. satellite imagery. Similarly with 2D projections, 1D relationships |, |∗,],]∗, [, [∗ and = are not present at all on x or y axes. This result confirms that adjacency relationship is not noticeable in DB, and it also shows that 2D projections do not describe well this relationship, since they are not able to detect it here. The high frequency of 1D relationships $<$ and $< *$ partially confirms the high frequency of ”disjoint” in the topological approach, and of areas ⊞ ⊞ ⊞ ⊞ with 9DSpa.

Among these three approaches, we think that 9DSpa is the one that allows providing the most relevant statistical knowledge for future interpretations. In particular, it is possible to deduce from them the probability of presence of a given entity in an area having a given directional relationship with a reference entity, as well as an indication on its size. For example, two entities of categories CHIMNEY (reference entity) and ROOF obtain the three best probabilities of presence 0.10, 0.14 and 0.17 with respective areas ⊞, ⊞ and ⊞. During an object detection and localization task, this knowledge gives the possibility to constrain the search of the target object to priority searching areas in the image and to corresponding object's size, given a reference object. All the associated statistics are available on the website[2] of this work.

5 Conclusion

We have presented a statistical study on spatial relationships of categories of entities from a public database of annotated images. This study provides a cartography of the spatial relationships that can be encountered in a database of heterogeneous natural contents. We think that it could be integrated with benefit in a knowledge-based system dedicated to artificial vision and CBIR, in order to enrich the description of the visual content as well as to help to choose the most discriminant type of relationships for each use case. Here, we have focussed on the analysis of unary and binary relationships. Study on unary relationships highlights trends on location of categories of entities in the image. These measures allows to determine the probability of the presence of a category in a given area, and to perform spatial reasoning. In the same way, study on binary relationships allows deducing the probability of presence of a category in an area regarding the location of another reference category. In addition, it gives indications on the relevance of the tested representations of these relationships. Ternary spatial relationships were already studied, but because of space limitation, they are not included in the paper ; see technical report [2] or the website[2].

This work was done on a manually annotated database of one thousand images. Therefore, it is evident that these statistics will have to be confirmed or refined on other image databases of larger size. However from now, we think that these measures can help us, on the one hand, to better understand which kinds of spatial relationship should be employed for a given problem and how to model them. On the other hand, such statistics can help to start a knowledge base on these relationships, that can be applied quickly to some topical problems of artificial vision and CBIR such as object detection, recognition or retrieval in a collection.

Acknowledgment. This work is supported by the French ANR project DISCO (2008-2010) and the Intl. CAPES-COFECUB collaboration AMIB (2010-2013).

References

1. Egenhofer, M.J., Al-Taha, K.K.: Reasoning about gradual changes of topological relationships. In: Proc. of the International Conference GIS, pp. 196–219. Springer, Heidelberg (1992)
2. Hoàng, N.V., Gouet-Brunet, V., Rukoz, M.: A cartography of spatial relationships in a symbolic image database. Note de recherche du LAMSADE, Universite Paris-Dauphine (April 2011)
3. Huang, P., Lee, C.: Image Database Design Based on 9D-SPA Representation for Spatial Relations. TKDE 16(12), 1486–1496 (2004)
4. Mayhew, D.: Principles and guidelines in software user interface design. Prentice-Hall, Englewood Cliffs (1992)
5. Nabil, M., Shepherd, J., Ngu, A.H.H.: 2D projection interval relationships: A symbolic representation of spatial relationships. In: Symposium on Large Spatial Databases, pp. 292–309 (1995)
6. Park, J., Govindaraju, V., Srihari, S.N.: Genetic engineering of hierarchical fuzzy regional representations for handwritten character recognition. International Journal on Document Analysis and Recognition (2000)
7. Russell, B., Torralba, A., Murphy, K., Freeman, W.: Labelme: a database and web-based tool for image annotation. In Proc. of the International Journal of Computer Vision 177(1-3), 157–173 (2008)
8. Wang, W., Zhang, A., Song, Y.: Identification of objects from image regions. In: Int. Conf. on Multimedia and Expo. (2003)

Multi-class Object Detection with Hough Forests Using Local Histograms of Visual Words

Markus Mühling, Ralph Ewerth, Bing Shi, and Bernd Freisleben

Department of Mathematics & Computer Science, University of Marburg,
Hans-Meerwein-Str. 3, D-35032 Marburg, Germany
{muehling,ewerth,shib,freisleb}@informatik.uni-marburg.de

Abstract. Multi-class object detection is a promising approach for reducing the processing time of object recognition tasks. Recently, random Hough forests have been successfully used for single-class object detection. In this paper, we present an extension of random Hough forests for the purpose of multi-class object detection and propose local histograms of visual words as appropriate features. Experimental results for the Caltech-101 test set demonstrate that the performance of the proposed approach is almost as good as the performance of a single-class object detector, even when detecting a large number of 24 object classes at a time.

Keywords: Multi-class object detection, object recognition, Hough forests.

1 Introduction

The task of finding a given object category in an image or video sequence has received considerable attention in the literature. While early approaches were sensitive to real world imaging conditions such as pose and occlusion, significant progress has been made in recent years [4]. In general, the task of object detection is posed as a binary classification problem. Object models are learned to distinguish between specific object classes and background. To detect multiple object classes, the usual procedure is to use a large number of independently trained single-class object detectors. However, this approach is computationally expensive and does not scale to thousands of object classes. Moreover, in the field of semantic concept detection it has been shown that it is beneficial to integrate object detection results as additional inputs for a support vector machine (SVM) [13]. For this purpose, many object classes have to be detected in large image and video databases. An appealing approach to reduce the computational overhead is the concurrent detection of several object classes by using a multi-class learning framework. Instead of learning class-specific object detectors, the aim is to learn a common classification model for multiple or all object classes.

In this paper, we present an extension of Hough forests [7] for multi-class object detection. Gall and Lempitsky's Hough forests [7] are random forests that use local features to vote for object locations in order to realize a generalized

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 386–393, 2011.

Hough transform for a single-class object detection problem. We show that the features and the split function used in Gall and Lempitsky's approach are not appropriate for multi-class object detection. To resolve this issue, local histograms of visual words (HoW) in conjunction with an appropriate node split function for multi-class random Hough forests are proposed. The presented approach can classify multiple classes without any significant computational overhead, in contrast to other multi-class approaches such as multi-class SVMs that have to build one-against-one classifiers for all class combinations. Experimental results for the Caltech-101 test set demonstrate that the presented multi-class approach relying on local HoWs achieves similar performance as (single) class-specific Hough forests, even when detecting as many as 24 object classes at a time.

The remainder of the paper is organized as follows. Related work is discussed in Section 2. In Section 3, the construction of multi-class Hough forests is explained. In Section 4, experimental results are presented. Section 5 concludes the paper and outlines areas for future research.

2 Related Work

Random forests introduced by Breiman [3] consist of an ensemble of decision trees. They inherit the positive characteristics of decision trees, but they do not suffer from the problem of overfitting. Breiman has shown empirically that random forests are more robust to noise in the training data, i.e., mislabeled training examples, than Adaboost. Moreover, the construction of trees has a rather low computational complexity compared to a SVM, classification is very efficient at runtime, and training as well as classification can be easily parallelized at the level of decision trees. Recently, random forests have been successfully applied to image classification. Bosch et al. [1] have investigated random forests for multi-class image classification using spatial shape and appearance descriptors. Simple linear classifiers on random feature subsets are used as decision functions within the trees. The authors have shown that random forests are significantly faster with only a slight performance decrease than a multi-class multiple kernel SVM.

In the field of object detection, Hough-based approaches have achieved impressive detection performance and speed. Exploiting the fact that object parts provide useful spatial information, local features are used to vote for object locations. Thus, these approaches are relatively robust to partial occlusions, shape and appearance variations. Leibe et al. [10] presented the implicit shape model as a probabilistic formulation of the Hough transform. It consists of a class-specific codebook and a spatial probability distribution. The codebook is learned from local feature descriptors using the k-means clustering algorithm, and the probability distribution specifies where each codebook entry can be found within the object area. At the detection stage, local descriptors are matched to codebook entries, and probabilistic Hough votes are generated based on the corresponding spatial probability distribution. The Hough votes are aggregated in a voting space where object locations are determined by searching for local maxima. To optimize the detection performance, Maji and Malik [12] have extended the implicit shape model by placing the Hough transform in a discriminative framework. The authors use a max-margin formulation to learn weights on the codebook entries. The weights for possible object location votes indicate whether a codebook entry is a good predictor for an object location. Another way of improving object detection performance is to discriminatively learn the codebook. Gall and Lempitsky [7] use the random forest framework, called Hough forest, to realize a generalized Hough transform to detect object appearances. Kumar and Patras [9] use a different criterion based on intermediate Hough images for tree construction. They try to explicitly maximize the response at the true object locations in the Hough images. Therefore, Hough spaces for all training images have to be calculated at all non-leaf nodes during training. Hough forests are also used by Fanelli et al. [5] for mouth localization in facial images and by Yao et al. [16] for action recognition in videos. An approach for multi-class object detection has been presented by Torralba et al. [14]. Instead of training object detectors individually, the authors use a joint-boosting algorithm to share features among object classes. Using 21 object classes from the LabelMe dataset, the authors have shown that jointly learning object classes need less training data and yield a better object detection performance than single object detectors.

3 Multi-class Hough Forests

The proposed multi-class object detection approach is based on the class-specific Hough forest presented by Gall and Lempitsky [7]. We extend this approach for multi-class object detection. Furthermore, different local feature representations as described in Section 3.1 are investigated. The construction of the underlying random forest, including the Hough voting extension and the required leaf node information, is explained in Section 3.2. Finally, the Hough voting process and the detection of object centers is presented in Section 3.3.

3.1 Local Features

The random Hough forest proposed by Gall and Lempitsky [7] uses decision functions that directly compare pixel values. Therefore, each local patch consists of a number of image channels: three color channels, four edge channels with first- and second-order derivatives and nine HOG-like (histogram of gradients) channels. Apart from these HOG-like features, we investigate two further feature representations. First, densely sampled SIFT (Scale Invariant Feature Transform [11]) descriptors are used. To extract these features, the Vision Lab Features Library (VLFEAT) [15] is used, because it provides a fast algorithm for the calculation of a large number of SIFT descriptors of densely sampled features of the same scale and orientation. Color information is integrated by concatenating and computing the descriptors independently for the three channels of the RGB color model. Due to the normalizations, the RGB-SIFT descriptor is invariant against light intensity and color changes or shifts, respectively.

Furthermore, based on the previously described SIFT descriptors, the usefulness of local HoWs is analyzed. We assume that these descriptors are more suitable to describe local object parts, because they capture the local spatial arrangement of visual words. In a first step, a vocabulary of visual words is generated by clustering the SIFT descriptors from the set of training images in their feature space. For this purpose, the k-means algorithm is used to derive k cluster centers that represent the visual words. A vocabulary size of 1000 visual words is used in our experiments. Then, image regions are represented as local HoWs by mapping the keypoint descriptors of a local region to the visual words of the vocabulary. During histogram generation, the similarity of keypoint descriptors and vocabulary entries is calculated according to the soft-weighting scheme of Jiang et al. [8]: Instead of mapping a keypoint only to its nearest neighbor, the top K nearest visual words are selected. Using a visual vocabulary of N visual words, the importance of a visual word t in the image is represented by the weights of the resulting histogram bins $w = [w_1, \ldots, w_t, \ldots, w_N]$ with

$$w_t = \sum_{i=1}^{K} \sum_{j=1}^{M_i} \frac{1}{2^{i-1}} sim(j,t) \quad \text{and} \quad sim(x,y) = e^{-\frac{2}{\gamma} d(x,y)} \tag{1}$$

where M_i is the number of keypoints whose i-th nearest neighbor is the visual word t, d is the Euclidean distance and γ is the maximum distance between two codebook entries.

3.2 Random Forest Construction

A random forest is an ensemble of decision trees. To realize efficient multi-class object detection, the decision trees are trained in a multi-class fashion. The training data consist of a set of image patches $P_i = (I_i, c_i, d_i)$ with the class label c_i and the vector d_i describes the relative position to the object center. The training subsets for the different trees are generated using subbagging. The decision trees are built in a top-down manner by selecting at each node the best split function of a set of randomly instantiated split functions, so that the impurity of class labels and class specific offsets in the child nodes are minimized. Thus, to build the trees, a binary split function for decision making and an uncertainty measure have to be defined that guarantee the purity of class labels and offsets in the leaf nodes.

Split Function. Two different split functions are used. The decision function of the original approach of Gall and Lempitsky directly compares values of a pair of pixels in an image patch I within the same channel a:

$$t_{a,p,q,r,s,\tau}(I) = \begin{cases} 0, & \text{if } I^a(p,q) < I^a(r,s) + \tau \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

with a decision threshold τ and two locations within the image patch (p, q) and (r, s). For local HoWs as well as for the edge histograms of SIFT descriptors, the following simple linear classifiers are applied:

$$t_{n,b}(x) = \begin{cases} 0, & \text{if } n^T x + b \leq 0 \\ 1, & \text{otherwise} \end{cases} \tag{3}$$

where n is a vector of the same size as the feature vectors. Randomness is introduced by randomly choosing the channel and pixel positions and in the case of linear classifiers by randomly choosing the components of the vector n in the range of $[-1, 1]$.

Uncertainty Measure. For each node, a set of decision functions t^k with randomly chosen parameters is considered. The following optimization function with $U \in \{U_1, U_2\}$ is solved to find the binary test that optimally splits the data:

$$argmin_k(U(P_i|t^k(I_i) = 0) + U(P_i|t^k(I_i) = 1). \tag{4}$$

The class-label uncertainty is modified for the multi-class case by

$$U_1(A) = |A| \cdot Entropy(A) \quad \text{with} \quad Entropy(A) = -\sum_{c=0}^{C-1} \frac{|A^c|}{|A|} \log_2 \left(\frac{|A^c|}{|A|} \right) \tag{5}$$

and the offset uncertainty by

$$U_2(A) = \sum_{c=1}^{C-1} \sum_{i:c_i=c} ||d_i - d_A^c||^2 \quad \text{with} \quad d_A^c = \frac{1}{|A^c|} \sum_{i:c_i=c} d_i \tag{6}$$

where C is the number of classes, A^c is the subset of A containing all instances of class c and d_i is the offset of the i-th local patch. For calculating the offset uncertainty, the background class is not considered. The type of uncertainty is randomly chosen for each node.

Leaf Node Information. The training data are recursively split until a maximum depth is reached or the number of patches falls below a minimum. The final leaf nodes represent the visual codebook and store the class as well as the spatial information. Each leaf node consists of a list of offset vectors and corresponding class labels for the containing instances. Furthermore, the class probabilities, i.e., the percentage of the corresponding object class patches, are stored. These probabilities later determine the weight of the associated Hough votes.

3.3 Hough Voting and Local Maxima Detection

During object detection, the local feature descriptors are propagated through the trees of the random forest according to the split criteria in the nodes. At the leaf nodes, Hough votes for locations of possible object centers are triggered using the stored offset vectors. The votes are weighted by the corresponding class probabilities. Two voting strategies are investigated. The first strategy votes for all classes in the leaf node. Thus, weighted votes are generated for all offset vectors. The second strategy only considers offset vectors from the dominating object class. To detect objects at different sizes, the Hough Forest is applied to a series of images at different scales resulting in several Hough images, one Hough image per object class and scale. Finally, the objects are detected as maxima in the Hough images. A Hough image contains the accumulated votes.

The idea of the Hough transformation is that the triggered votes of local patches yield peaks in the Hough image at the positions of the object centers. These local maxima in the Hough images are detected using the mean-shift algorithm, which is a local, iterative and non-parametric approach. The implementation of Intel's OpenCV library [2] is used. It uses a local search window of predefined size. The local maxima have to exceed a predefined threshold to be accepted as an object center. The bounding boxes are determined based on the scale of the corresponding Hough image.

4 Experiments

In this section, experimental results are presented for a subset of the Caltech-101 object test set [6]. Caltech-101 is a challenging dataset containing 101 object classes and a background class. The bounding boxes are provided as ground truth for all object appearances. For our experiments, we have randomly selected 24 object classes: "airplane", "bonsai", "brain", "buddha", "butterfly", "car", "chandelier", "ewer", "face", "grand piano", "hawksbill", "helicopter", "kangaroo", "ketch", "laptop", "leopard", "menorah", "motorbike", "revolver", "scorpion", "starfish", "sunflower", "trilobite", and "watch". For each object class, 65 randomly chosen images including background were used for training, and from the remaining images, 15 images were randomly chosen for each class for testing. The F1-score is calculated for the point in the ROC-curve where the difference of recall and precision is minimal. In a first experiment, the multi-class and single-class object detection performance on the 24-class subset of the Caltech-101 dataset is investigated. The experimental results are displayed in Table 1. The application of a large number of class-specific object detectors achieved better performances than the multi-class extensions. While the accuracy of the multi-class extension of the original approach declined from 56.7% to 14.4%, the approaches based on SIFT and local HoWs showed a significantly smaller performance decrease from 42.4% to 36.5% and from 57.9% to 54.5%, respectively. Overall, the best performance was achieved using HoWs. Furthermore, the experiments have been repeated with pre-scaled test images. The test images have been scaled such that the objects are of the same size as in the training set. The results for this experiment are also presented in Table 1. As expected, the performance increased for all runs. The implementation of Gall and Lempitsky's approach seems to be more sensitive to differing object scales than the proposed approach that relies on local HoW features. While the performance loss of the original approach comparing multi-class and single-class detection was 42.3% in the preceding experiment, it also declined by 31.4% using equally scaled objects. The performance loss of the proposed approach using HoW features amounts to only 2.8% when pre-scaled images are used. The experiments suggest that the combination of local HoW features and linear classifiers as decision functions is more appropriate for Hough forests for multi-class object detection. The multi-class object detectors are significantly faster than the single-class detectors (see Table 2). To further reduce the computation times the training of Hough forests

Table 1. Mean f1-scores for 24 classes of Caltech-101

In [%]			Pre-scaled test images	
	Single-class	Multi-class	Single-class	Multi-class
G&L	56.7	14.4	70.6	39.2
DSIFT	42.4	36.5	55.6	49.1
HOW	57.9	54.5	67.8	65.0

Table 2. Runtimes for 24 classes of Caltech-101 on a 2 GHz AMD Opteron^{TM} processor 270 with 8 GB RAM, running a Debian Linux 5.0.3, implemented in C++

	Training Runtime		Testing Runtime (per image)	
	Single-class	Multi-class	Single-class	Multi-class
DSIFT	1473 h	75 h	17.8 sec	0.9 sec
HOW	1284 h	67 h	74.6 sec	4.1 sec

is easily parallelizable on the tree level. Moreover, if object detection results are used for concept detection, the overhead for computing the HoWs is negligible since related state-of-the-art systems rely on visual words and thus these features do not need to be calculated twice.

5 Conclusions

To detect a large set of object classes in images, it is inefficient to run a large number of single-class object detectors. In this paper, we have presented a multi-class approach for the task of object detection. The presented approach is capable of detecting 24 different object classes at a time, instead of applying one object detector for each object class separately. To achieve this, we have extended a random Hough forest approach with appropriate measures for class and offset uncertainty. The proposed approach relies on local HoW features with an adequate split function. It turned out that the choice of features is crucial for obtaining a multi-class detection performance that is comparable to the single-class case. While the performance of the multi-class extension of the original approach using HoG-like features clearly dropped, the multi-class Hough forest based on local HoW features almost retained the performance compared to the class-specific version. Overall, it was shown how multi-class Hough forests can be constructed to speed up the concurrent detection of many object classes in images. Areas for future work are, for example, the investigation of multi-class detectors for different object subsets or the use of context information.

Acknowledgements. This work is supported by the German Ministry of Education and Research (BMBF, D-Grid) and by the German Research Foundation (DFG, PAK 509).

References

1. Bosch, A., Zisserman, A., Muoz, X.: Image Classification using Random Forests and Ferns. In: Proc. of the IEEE Int. Conf. on Computer Vision, pp. 1–8 (2007)
2. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)
3. Breiman, L.: Random Forests. Machine Learning 45, 5–32 (2001)
4. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision 88(2), 303–338 (2010)
5. Fanelli, G., Gall, J., Van Gool, L.: Hough Transform-based Mouth Localization for Audio-Visual Speech Recognition. In: Proc. of the British Mach. Vis. Conf. (2009)
6. Fei-Fei, L., Fergus, R., Perona, P.: Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. Computer Vision and Image Understanding 106(1), 59–70 (2007)
7. Gall, J., Lempitsky, V.: Class-Specific Hough Forests for Object Detection. In: Proc. of the IEEE Conf. on Comp. Vis. and Pat. Recog., pp. 1022–1029 (2009)
8. Jiang, Y.G., Ngo, C.W., Yang, J.: Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval. In: Proc. of the ACM Int. Conference on Image and Video Retrieval, pp. 494–501 (2007)
9. Kumar, V., Patras, I.: A Discriminative Voting Scheme for Object Detection using Hough Forests. In: Proc. of the British Machine Vision Conference Postgraduate Workshop, pp. 1–10 (2010)
10. Leibe, B., Leonardis, A., Schiele, B.: Robust Object Detection with Interleaved Categorization and Segmentation. Int. J. of Comp. Vis. 77(1-3), 259–289 (2008)
11. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
12. Maji, S., Malik, J.: Object Detection Using a Max-Margin Hough Transform. In: Proc. of the IEEE Conf. on Comp. Vis. and Pattern Recog., pp. 1038–1045 (2009)
13. Mühling, M., Ewerth, R., Freisleben, B.: Improving Semantic Video Retrieval via Object-Based Features. In: Proc. of the IEEE Int. Conference on Semantic Computing, pp. 109–115 (2009)
14. Torralba, A., Murphy, K.P., Freeman, W.T.: Sharing Visual Features for Multiclass and Multiview Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(5), 854–869 (2007)
15. Vedaldi, A., Fulkerson, B.: VLFeat: An Open and Portable Library of Computer Vision Algorithms (2008), http://www.vlfeat.org/
16. Yao, A., Gall, J., Van Gool, L.: A Hough Transform-Based Voting Framework for Action Recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2061–2068 (2010)

Graph Clustering Using the Jensen-Shannon Kernel

Lu Bai and Edwin R. Hancock*

Department of Computer Science, University of York
{lu,erh}@cs.york.ac.uk

Abstract. This paper investigates whether the Jensen-Shannon divergence can be used as a means of establishing a graph kernel for graph classification. The Jensen-Shannon kernel is nonextensive information theoretic kernel which is derived from mutual information theory, and is defined on probability distributions. We use the von-Neumann entropy to calculate the elements of the Jensen-Shannon graph kernel and use the kernel matrix for graph classification. We use kernel principle components analysis (kPCA) to embed graphs into a feature space. Experimental results reveal the method gives good classification results on graphs extracted from an object recognition database.

Keywords: Jensen-Shannon kernel, Divergence, Entropy, Graph Kernel.

1 Introduction

Graph based representations have proved to be powerful tools for structural pattern analysis in computer vision. One of the problems that arises with large amounts of graph data is that of edit distance computation which itself is dependant on accurate correspondence analysis.

There have been several successful attempts to classify graph data using clustering techniques. The methods developed include a) using vectors of structural characteristics or permutation invariants [18], b) applying pairwise clustering to edit distances [17], c) embedding graph structures in dissimilarity-based feature spaces, and d) through central clustering based on class prototypes [16] [15]. An alternative to this methods is to use kernel methods formulated in the graph domain [13]. Graph kernels aim to overcome the computational bottleneck associated with graph similarity or edit distance computation which is known to be NP-complete. This is analogous to the use of kernel methods with vector data, which allow efficient algorithms to be developed that can deal with high dimensional data without the need to construct an explicit high dimensional feature space [2]. A number of graph kernels have been reported in the literature and successfully applied to real world data [4]. Most of the reported methods share

* Edwin Hancock is supported by the EU FET project SIMBAD and by a Royal Society Wolfson Research Merit Award.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 394–401, 2011.

the feature of exploiting topological information concerning the arrangement of nodes and edges in a graphs. There are three popular methods, namely a) diffusion kernels defined with respect to similarity [9] [10], b) random walk kernels based on counting the number of nodes with the same label in a random walk [8] [4], and c) the shortest path kernel [1].

Recently, information theory has been used to define a new family of kernels on probability distributions, and these have been applied to structured data [13] [14] [3] [6]. These so-called nonextensive information theoretic kernels are derived from the mutual information between probability distributions, and are related to the Shannon entropy. Examples include the Jensen-Shannon kernel [12].

One of the problems in constructing Jensen-Shannon kernels for graphs is that of constructing the required probability distributions or computing the entropy associated with graph structures. Both of these problems have proved elusive, and this is turn has provided an obstacle to the successful construction of information theoretic graph kernels [12]. Recently, we have shown to efficiently compute the von Neumann entropy of a graph [11]. The von Neumann entropy is the Shannon entropy associated with the Laplacian eigenvalues of a graph, and requires the computation of the graph spectrum, and this is cubic in the number of nodes. By approximating the Shannon entropy by its quadratic counterpart, we have shown how to the computation can be rendered quadratic in the number of nodes. In this paper, we explore how this simplification can be used to efficiently compute the Jensen-Shannon kernel between graphs. The resulting computations depend on the node degree distribution over the graph and can be simply computed for both the original graphs and their tensor product. Once the Jensen-Shannon kernel is to hand, we use kernel principle components analysis (kPCA) [7] to embed the graphs into a low dimensional feature space where classification is performed.

This paper is organised as follows. Section 2 briefly reviews the basic concepts of Jensen-Shannon kernel. Section 3 reviews how we construct a node degree probability distribution over both graphs and product graphs. This distribution is used to approximate the von-Neumann entropy, and we show how to calculate Jensen-Shannon graph kernel. Section 4 provides our experimental evaluation. Finally, Section 5 provides conclusions and directions for future work.

2 Jensen-Shannon Kernel

In this section we review the basic theory of the Jensen-Shannon Kernel used in our work. The Jensen-Shannon Kernel is a nonextensive mutual information kernel which is defined using the extensive and nonextensive entropy. It is defined on probability distributions over structured data. Assume $M^1_+(\chi)$ is a set of probability distributions where χ is a set provided with some $\sigma - algebra$ of measurable subsets, the kernel $k_{JS} : M^1_+(\chi) \times M^1_+(\chi) \to R$, is positive definite (pd) with the following kernel function [12]:

$$k_{JS}(P_1, P_2) = \ln 2 - JS(P_1, P_2) \tag{1}$$

where $JS(P_1, P_2)$ is the Jensen-Shannon divergence $k_{JS} : M^1_+(\chi) \times M^1_+(\chi) \rightarrow [0, \infty)$defined as

$$JSD(P_1, P_2) = H(\frac{P_1 + P_2}{2}) - \frac{1}{2}(H(P_1) + H(P_2)) \tag{2}$$

and for a mixture of n probability distribution $P_1, ...P_n$ with mixing proportions $\pi_1, ..., \pi_n$, the divergence is given by:

$$JSD(P_1, P_2, ..., P_n) = H(\Sigma^n_{i=1}\pi_i P_i) - \Sigma^n_{i=1}\pi_i H(P_i) \tag{3}$$

where $H(P_i)$ is an Shannon entropy for distribution P_i.

3 Jensen-Shannon Graph Kernel

In this section we explore how to compute the Jensen-Shannon kernel for pairs of graphs. We commence by defining a probability distribution over the node degree distribution and then show how this can be used to compute the Shannon entropy appearing in the definition of the kernel.

3.1 Node Degree Distribution

We use the node degree distribution to calculate the Jensen-Shannon graph kernel. To commence, we denote the graph as $G = (V, E)$ where V is the set of nodes and $E \subseteq V \times V$ is the set of edges. The adjacency matrix A of graph G has elements

$$A(u, v) = \begin{cases} 1 & if(u, v) \in E, \\ 0 & otherwise. \end{cases} \tag{4}$$

The degree matrix of graph G is a diagonal matrix D with the nodes degrees as diagonal elements $D(u, u) = d_u = \sum_{u,v \in V} A(u, v)$. From the adjacency matrix and the degree matrix we compute the Laplacian matrix $L \equiv D - A$, which has elements

$$L(u, v) = \begin{cases} d_v & if u = v, \\ -1 & if(u, v) \in E, \\ 0, & otherwise. \end{cases} \tag{5}$$

The spectral decomposition of the Laplacian matrix is $L = \sum_{u=1}^{|V|} \lambda_i \phi_i \phi_i^T$ where λ_i are the eignevalues and ϕ_i are the eigenvectros of L.

We can define the node degree distribution as the node degree divided by the volume of the graph, and for node $u \in V$ the probability is

$$P_G(u) = d_u / \sum_{v \in V} d_v \tag{6}$$

3.2 Graph Product

Before we introduce the Jensen-Shannon graph kernel, we first introduce the graph product concept. For the graphs $G(V, E)$ and $G'(V', E')$ the product graph $G\times = (V\times, E\times)$, has node and edge sets

$$V_{\times} = \{(v_i, v_i^{'}) : v_i \in V \wedge v_i^{'} \in V\} \tag{7}$$

$$E_{\times} = \{((v_i, v_i^{'}), (v_i, v_i^{'})) : (v_i, v_i^{'}) \in E \wedge (v_i, v_i^{'}) \in E^{'}\} \tag{8}$$

If A and A' are the adjacency matrices of graphs G and G' respectively $A\times = A \prod A'$ is the adjacency matrix of the product graph $G\times$. The most common graph products are formed by taking the Cartesian product, tensor product or the union. For reasons of efficiency here we take the union graph. We compute the difference in entropy between the two graphs and their union. To construct the union graph, we perform pairwise correspondence matching. Details of the construction are outside the scope of this paper. Our approach follows that of Lin, Wilson and Hancock [5], and the adjacency matrix of the union is denoted by A_U.

3.3 Jensen-Shannon Graph Kernel Graph Kernel

Consider a a graph set $\{G_1, ..., G_i, ..., G_j, ..., G_n\}$. A graph kernel can be defined using a similarity measure to compute the $n \times n$ positive matrix. Associated with the degree distribution P_i and P_j of graphs G_i and G_j, the Jensen-Shannon kernel is defined as

$$k_{JS}(P_i, P_j) = \ln 2 - H(\frac{P_i + P_j}{2}) + \frac{1}{2}(H(P_i) + H(P_j)) \tag{9}$$

We suppose $\frac{P_i+P_j}{2}$ represents the degree distribution of the product graph $G_{\times}$ of $G_i(V_i, E_i)$ and $G_j(V_j, E_j)$. As a result (9) can be written as

$$k_{JS}(P_i, P_j) = \ln 2 - H(P_{\times}) + \frac{1}{2}(H(P_i) + H(P_j)) \tag{10}$$

Here we use the von Neumann entropy to compute the Jensen-Shannon kernel. The von Neumann entropy for graph G is $H_{VN} = \sum_{i=1}^{|V|} \frac{\lambda_i}{2} ln \frac{\lambda_i}{2}$. By approximating the non-Neumann entrpy by its quadratic counterpart, Han and Hancock [11] have shown that the approximate von-Neumann entropy is given by

$$H_{VN} = \frac{|V|}{4} - \sum_{(u,v)\in E} \frac{1}{4 d_u d_v} \tag{11}$$

As the node degree distribution $P_i(u)$ and $P_j(v)$ can be written as $P_i(u) = d_u / \Sigma_u^V d_u$ and $P_j(v) = d_v / \Sigma_v^V d_v$, so associated with function (10) and (11), the Jensen-Shannon graph kernel can be approximated as

$$k_{JS}(P_i,P_j) = \ln 2 - (\frac{|V_\times|}{4} - \sum_{v_\times \neq u_\times} \frac{1}{P_\times(u_\times)P_\times(v_\times)}) + \frac{|V_i|+|V_j|}{2}$$
$$- \frac{1}{2}(\sum_{u_1\neq v_1, v_1(i)\neq u_1(j)}^{V_1} \frac{1}{P_i(u)P_i(v)} + \sum_{u_2\neq v_2, v_2(i)\neq u_2(j)}^{V_2} \frac{1}{P_j(u)P_j(v)}) \tag{12}$$

where $P_\times(i)$ represents the degree distribution of $p_\times$.

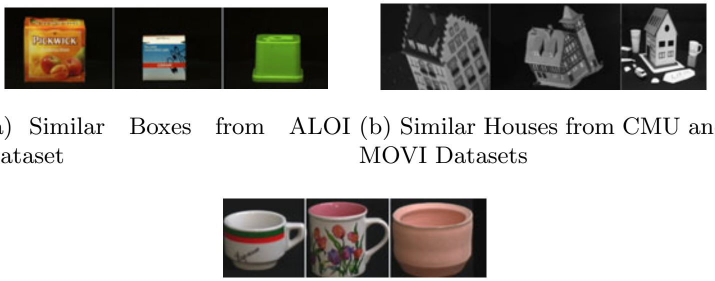
4 Experiments

In this section, we will explore whether the Jensen-Shannon kernel can be used for object recognition, and evaluate its stability. we commence by classifying synthetic graph abstracted from real-world image, and then calculate relationship between the kernel value and number of edit operation.

4.1 Graph Characterization

In the first experiment, we use three different graph datasets to illustrate the classification performance of the Jensen-Shannon graph kernel. The first dataset consists of graphs are extracted from digital images of three similar boxes in the ALOI database Fig.1(a), the second of graphs extracted from images of three toy houses in the MOVI and CMU databases, and the third of graphs extracted from images of cups from the COIL database Fig.1(3). For each object there are 18 images captured from different viewpoints. The graphs are the Delaunay triangulations of feature points extracted from the different images.

For the three different datasets we compute the Jensen-Shannon kernel matrices. We perform kernel Principal Components Analysis (kPCA) on the kernel matrices to embed the graphs into a 3-dimensional feature space. Fig.2(1),

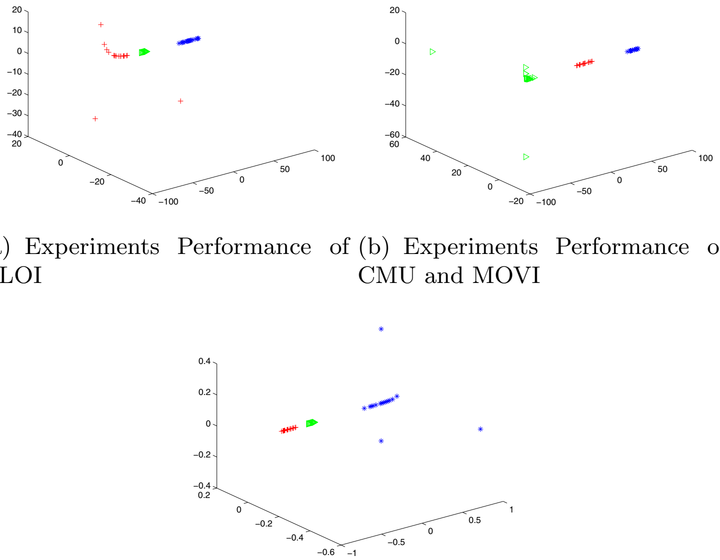


(a) Similar Boxes from ALOI Dataset

(b) Similar Houses from CMU and MOVI Datasets

(c) Similar Cups from COIL Datasets

Fig. 1. Datastes for Experiments



(a) Experiments Performance of ALOI

(b) Experiments Performance of CMU and MOVI

(c) Experiments Performance of COIL

Fig. 2. Experiment Performance of Jensen-Shannon Kernel

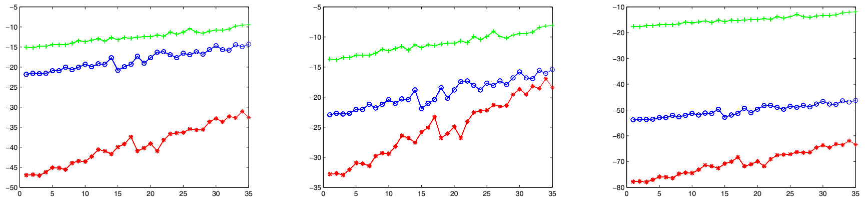
Table 1. Accurancy of Classification with Jensen-Shannon Kernel

Datasets	Object1	Object2	Object3
Boxes (ALOI)	100%	100%	100%
Houses (CMU and MOVI)	100%	100%	100%
Cups (COIL)	100%	100%	100%

Fig.2(b) and Fig.2(c) show the resulting embeddings. In each case the different objects are well separated in the embedding space. To place our analysis on a more quantitative footing, we apply K-means clustering method to the embedded graphs, and compute the classification accuracy for the three datasets. Table.1 summaries the results, and indicate that an accuracy of 100% is achievable.

4.2 Stability Evaluation

Next we evaluate the stability of the Jensen-shannon kernel. We select nine seed graphs from the three groups of real-world images shown in Fig.1. We then apply random edit operations to the nine seed graphs to simulate the effects of noise. The edit operations are node deletion and edge deletion. For each seed graph, we randomly delete a predetermined fractions of the nodes or edges to obtain noise corrupted variants. Fig.3. and Fig.4. show the effects of node deletion and

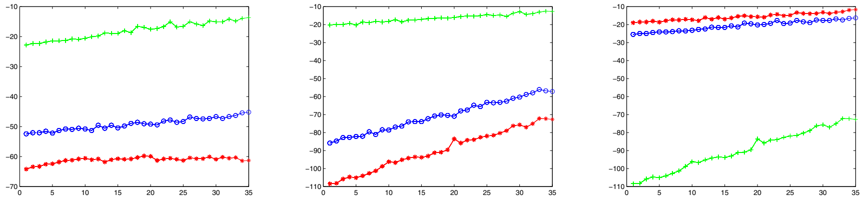


(a) Evaluation of ALOI (node)

(b) Evaluation of CMU and MOVI (node)

(c) Evaluation of COIL (node)

Fig. 3. Jensen-Shannon Kernel Evaluation with Node Deletion Edit Operation



(a) Evaluation of ALOI (edge)

(b) Evaluation of CMU and MOVI (edge)

(c) Evaluation of COIL (edge)

Fig. 4. Jensen-Shannon Kernel Evaluation with Edge Deletion Edit Operation

edge deletion for each group of graphs respectively. The x-axis shows the fraction of nodes or edges deleted, and the y-axis shows value of the kernel $K(G_o, G_n)$ between the original graph G_o and its noise corrupted counterpart G_n. The plots show that there is an approximately liner relationship between the Jensen-Shannon kernel and the number of deleted nodes or edges, i.e. the graph edit distance.

5 Conclusions

In this paper, we have shown how to construct Jensen-Shannon kernels for graph data-sets using the von-Neumann entropy. The method is based on a probability distribution over the node degree in a graph, and uses the von Neumann entropy to measure the mutual information between pairs of graphs. By applying kernel PCA to the Jensen-Shannon kernel matrix, we embed sets of graphs into a low dimensional space. Here we use K-means clustering to assign the graphs to classes. Experimental results reveal that the method gives good results for graph datasets extracted from image data.

References

1. Borgwardt, K.M., Kriegel, H.P.: Shortest-path kernels on graphs (2005)
2. Bunke, H., Riesen, K.: Graph classification based on dissimilarity space embedding. In: Structural, Syntactic, and Statistical Pattern Recognition, pp. 996–1007 (2010)
3. Desobry, F., Davy, M., Fitzgerald, W.J.: Density kernels on unordered sets for kernel-based signal processing. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, vol. 2, p. II–417. IEEE, Los Alamitos (2007)
4. Gartner, T.: A survey of kernels for structured data. ACM SIGKDD Explorations Newsletter 5(1), 49–58 (2003)
5. Han, L., Hancock, E., Wilson, R.: Learning generative graph prototypes using simplified von neumann entropy. In: Graph-Based Representations in Pattern Recognition, pp. 42–51 (2011)
6. Jebara, T., Kondor, R., Howard, A.: Probability product kernels. The Journal of Machine Learning Research 5, 819–844 (2004)
7. Jolliffe, I.: Principal component analysis (2002)
8. Kashima, H., Tsuda, K., Inokuchi, A.: Marginalized kernels between labeled graphs. In: International Workshop then Conference on Machine Learning, vol. 20, p. 321 (2003)
9. Kondor, R.I., Lafferty, J.: Diffusion kernels on graphs and other discrete input spaces. In: International Workshop then Conference on Machine Learning, pp. 315–322. Citeseer (2002)
10. Lafferty, J., Lebanon, G.: Diffusion kernels on statistical manifolds. The Journal of Machine Learning Research 6, 129–163 (2005)
11. Lin, H., Hancock, E.R.: Characterizing Graphs Using Approximate von-Neumann Entropy (2011)
12. Martins, A.F.T., Smith, N.A., Xing, E.P., Aguiar, P.M.Q., Figueiredo, M.A.T.: Nonextensive information theoretic kernels on measures. The Journal of Machine Learning Research 10, 935–975 (2009)
13. Scholkopf, B., Smola, A.J.: Learning with kernels. Citeseer (2002)
14. Shawe-Taylor, J., Cristianini, N.: Kernel methods for pattern analysis. Cambridge Univ Pr., Cambridge (2004)
15. Suau, P., Escolano, F.: Bayesian optimization of the scale saliency filter. Image and Vision Computing 26(9), 1207–1218 (2008)
16. Torsello, A., Hancock, E.R.: Graph embedding using tree edit-union. Pattern recognition 40(5), 1393–1405 (2007)
17. Torsello, A., Robles-Kelly, A., Hancock, E.R.: Discovering shape classes using tree edit-distance and pairwise clustering. International Journal of Computer Vision 72(3), 259–285 (2007)
18. Xiao, B., Hancock, E.R., Wilson, R.C.: Graph characteristics from the heat kernel trace. Pattern Recognition 42(11), 2589–2606 (2009)

CSS-AFFN: A Dataset Representation Model for Active Recognition Systems

Elizabeth González, V. Feliú, and A. Adán

E.T.S. Ingenieros Industriales, University of Castilla La Mancha, Avda. Camilo Jose Cela s/n,
Ciudad Real, 13071, Spain
Elisabeth.Gonzalez@uclm.es

Abstract. This paper proposes an object database representation model for active recognition systems. This model optimizes dataset information. The objects are modeled by using the proposed Canonical Sphere Section (CSS) model and the shape is normalized to affine transformations in the spectral dominium. This dataset representation model is compared with other shape representation models and implemented in an active recognition system which develops object manipulation. Its feasibility in complex robotic applications is therefore proved.

Keywords: Affine normalization, active object recognition, object modeling.

1 Introduction

The use of monocular vision to develop 3D object recognition provides several advantages to robotic applications: low economical cost of its implementation (in several cases the use of a webcam would be sufficient) and the fact that it can be easily integrated into industrial applications under real time requirements. However, the transformation of a 3D object recognition into a 2D shape recognition problem implies an undesirable factor: uncertainty. The uncertainty signifies that one specific view of an object might be similar to other objects views due to: the loss of loss of information (ambiguity) and factors such as noise, illumination variations, occlusions, etc. that may corrupt the feature vector describing the object view in the scene. The uncertainty problem can be solved by moving the sensor to other positions in order to collect more evidence about the scene object.

Several active recognition methodologies have been developed [1,2,3,4]. These recognition systems show high recognition rates even in the cases of objects with a high level of ambiguity, but their computational efficiency is low when the dataset have a large number of objects as a result of the large number of views that must be handled in the dataset and the number of robot movements required to identify an object in the scene. They are focused on solving the ambiguity/uncertainty problem without considering the object representation model. Furthermore, a large number of views representing the object model increase the number of items in the dataset, and the searching process to identify the scene view in the dataset therefore become slower. Moreover, in

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 402–409, 2011.

most cases researchers use a probabilistic framework to solve the uncertainty problem [2,3], which requires a shape descriptor with stochastic properties such as a PCA descriptor [5]. However, active recognition systems based on PCA have an important drawback: the object pose cannot be estimated with accuracy and their application to tasks like object manipulation it is not recommended. It is also advisable to choose other shape representation models to provide accuracy in object pose estimation. Although shape representation models are a widely researched subject [6], their implementation in active object recognition systems requires the satisfaction of several constraints such as: robustness to shape variations caused by viewpoint variation, scene noise, illumination changes, etc. Researchers deal with those problems using a feature vector invariant to affine transformations, but a critical problem concerns the number of elements in the feature vector that are necessary to describe, with a given precision, all the shapes in a database in which the number of different shapes is very high. From our experience we know that the recognition rates increase or decrease according to the number of elements in the feature vector and that, depending on the shape boundary, more or less elements must be used to describe it. If a new object is added to the database, most recognition systems therefore have to re-run several tests in order to find the optimal number of elements in the feature vector to achieve the best shape recognition rates.

The contribution of our work lies in defining an optimal database object representation model in an attempt to solve the two questions mentioned below: (1) which is the number of views needed to model the object and (2) a shape representation model that is invariant to affine transformations, able to represent all the shapes in the dataset with the same precision. We first propose a method with which to select the object canonical views according to object symmetry (Canonical Sphere Section). For the canonical views, we suggest a model to represent the views in the dataset in the spectral dominium. This representation will be referred to as the Affine Spectral Silhouette Normalized (ASSN) model, which is invariant to affine transformations and it is able to represent all the dataset shapes with the same precision using a feature vector of a fixed number of elements. This proposed database representation model have been implemented in an active recognition to evaluate its performance with regard to other active recognition systems from the scientific literature. Our dataset representation model has also been used in a robotic-vision system to develop object manipulation in order to show its feasibility in object manipulation, which is a complex robotic application. The following sections provide a detailed explanation of the database object representation model and its performance in a robotic-vision system.

2 Object Model Representation

Since the objects are represented by their appearance from a viewing sphere, it is advisable to select the sphere section whose viewpoints correspond with the *Canonical Object Section (COS)*. The sphere section containing the viewpoints associated with the *Canonical Object Section* will be denoted as *Canonical Sphere Section (CSS)*. Figure 1 shows samples of objects with different symmetries, their *Canonical Object Section* and their *Canonical Sphere Section*.

Defining the Canonical Spherical Section (CSS)

Let us take a synthetic object, o, whose principal axes have been aligned to the canonical coordinate system, as in Figure 2. Figure 2(a) shows a viewing sphere $\mathbb{S}$ represented in the canonical coordinate system, while, Figure 2(b) shows the principal bottle axes. Figure 2(c) shows the result of the alignment of principal object axes with the canonical coordinates system, in which we can see the object pose inside $\mathbb{S}$ where the object's major axis corresponds with axis Y and the minor axis corresponds with the X axis.

The o model is viewed from different view directions corresponding to the nodes of the sphere $\mathbb{S}$ with J nodes: this sphere has a radius r, and the image is projected in the direction defined by the node with the camera pointing towards the sphere center. These nodes are labeled by integer positive numbers j following a given order structure. Let us associate the view j of the object with a synthetic image $I_j, 1 \leq j \leq J$. The position of node j, which is representative of the view j, is given in spherical coordinates by a pair of angles (ψ_j, σ_j), $\psi_j \in [0 \;\; 2\pi]$ and $\sigma_j \in \left[\frac{\pi}{2} \;\; -\frac{\pi}{2}\right]$, these being the azimuth and polar coordinates respectively.

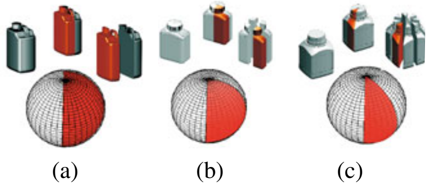
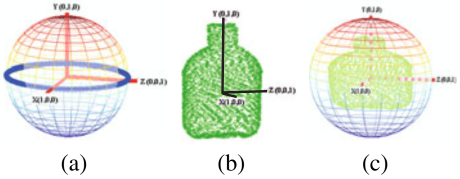

(a) (b) (c)

Fig. 1. Samples of objects with different symmetries, their *Canonical Object Section* and their *Canonical Sphere Section*



(a) (b) (c)

Fig. 2. a) Sphere coordinates system b) Object principal axes. c) Object principal axis aligned with the reference coordinate system.

Therefore, let n_z and n_y be the order of symmetry which is reflective for axes Z and Y respectively. The nodes j^* in the *Canonical Sphere Section* Ξ are defined as: $j^* = j \in \Xi : \psi_j \in [0 \;\; \frac{\pi}{n_y}], \sigma_j \in [0 \;\; \frac{\pi}{n_z}]$.

The relationship between the nodes in Ξ and the other nodes in $\mathbb{S}$ is: $I_j = I_{j^*} : \psi_j = (k_y \centerdot \frac{2\pi}{n_y} + \psi_{j^*}), \sigma_j = (k_z \centerdot \frac{2\pi}{n_z} + \sigma_{j^*})$, $I_j = \hat{I}_{j^*} : \psi_j = (k_y \centerdot \frac{2\pi}{n_y} - \psi_{j^*}), \sigma_j = (k_z \centerdot \frac{2\pi}{n_z} - \sigma_{j^*})$ where $j \neq j^*$, $\hat{I}_{j^*}$ is the image reflection, $k_z = 1, 2, ...,$ and $k_y = 1, 2, ...$

Based on the relationship between the nodes in Ξ and the other nodes in $\mathbb{S}$, it is simple to compute the order of the reflective symmetry n_y and n_z over axes Y and Z respectively. See in [7] how to compute n_y and n_z values.

3 ASSN Descriptor

The ASSN descriptor is developed by means of three operations: 1- Contour Normalization, 2- Number of descriptor reductions, 3- Precision normalization.

The contour normalization process normalizes the contour to rotation, translation scale and skew. The contour normalization method used to model the database canonical views in the spectral dominium is equivalent to that proposed by [8] in the spatial dominium. Working in the spectral dominium increases the robustness because in [8] the normalization process to the staring point (rotation) is unstable, while rotation invariance is guaranteed in the spectral dominium.

Assume a silhouette s composed of N points $(x(w), y(w)), 1 \leq w \leq N$ on the plane: XY where the origin of index w is an arbitrary point of the curve, and w and $w+1$ are consecutive points according to a given direction (for example clockwise direction) over the silhouette. In order to normalizes the silhouette s to affine variations (scale, rotation, translation and skew), a set of linear operations are applied to the silhouette representation in the spectral dominium based in contour orthogonalization from [8]. Thus, to the silhouette coordinates $x(w)$ and $y(w)$, the Fast Fourier Transform $(\mathscr{F})$ is applied obtaining $X(m) = \mathscr{F}(x(w))$ and $Y(m) = \mathscr{F}(y(w))$ where $1 \leq m \leq N$. Then, the follow operations are developed:

1. The center-of-gravity of the curve is normalized so as to coincide with the origin:

$$X_1(m) = X(m) - \mu_x, Y_1(m) = Y(m) - \mu_y \tag{1}$$

   where $\mu_x = X(0)/N$, $\mu_y = Y(0)/N$
2. The curve is scaled horizontally and vertically

$$X_2(m) = X_1(m) \cdot \rho_x \cdot N, Y_2(m) = Y_1(m) \cdot \rho_y \cdot N \tag{2}$$

   where $\rho_x = \frac{1}{\sqrt{\Sigma|(X_1(m))^2|}}$, $\rho_y = \frac{1}{\sqrt{\Sigma|(Y_1(m))^2|}}$
3. The curve is rotated $\pi/4$ counterclockwise

$$X_3(m) = \frac{1}{\sqrt{2}} \cdot (X_2(m) - Y_2(m)), Y_3(m) = \frac{1}{\sqrt{2}} \cdot (X_2(m) + Y_2(m)) \tag{3}$$

4. The curve is scaled horizontally and vertically

$$X_4(m) = X_3(m) \cdot \tau_x \cdot N, Y_4(m) = Y_3(m) \cdot \tau_y \cdot N \tag{4}$$

   where $\tau_x = \frac{1}{\sqrt{\Sigma|(X_3(m))^2|}}$, $\tau_y = \frac{1}{\sqrt{\Sigma|(Y_3(m))^2|}}$

Finally, the silhouettes feature vector is: $S(m) = X_4(m) + i \cdot Y_4(m)$, where $|S(m)|$ is invariant to rotation, translation, scale and skew transformations.

The other two operations are supported by the following property in the spectral dominium: most significant harmonics are in the first and last positions of the Fourier descriptor vector while non important contour details in the spatial dominium correspond to the central harmonics. What is more, the reduction of the number of descriptors is based on eliminating the central harmonics in order to achieve a new vector with the desirable length $(L, L < N)$. So, if $S = S(1), ..., S(N)$, the reduced vector is $\hat{S} = \{S(1), S(2), ..., S(L/2), S(N - L/2), S(N - L/2 + 1), ..., S(N)\}$ where L is an even number.

The precision representation normalization step sets the central harmonic to zero according to the spatial representation error between the initial silhouette and that which is normalized. Indeed, for a predefined precision error $(\mathscr{A})$, we have set to zero the maximal number of central harmonic keeping the error between both silhouette minor or equal to ($\mathscr{A}$). Thus,

$$\hat{\hat{S}} = \{\hat{S}(1),\hat{S}(2)...,\hat{S}(K^*),0...0,\hat{S}(L-K^*),\hat{S}(L-K^*+1),...\hat{S}(L)\} : \tag{5}$$
$$K^* = maxK, K \in \mathbb{N}, \{|(x(l)-x'(l)|,|y(l)-y'(l))|\} \leq \mathscr{A}, \forall 1 \leq l \leq L,$$
$$(x(l),y(l)) = \mathscr{F}^{-1}(\hat{S}), (x'(l),y'(l)) = \mathscr{F}^{-1}\{\hat{S}(1),\hat{S}(2)...,\hat{S}(K),0...0,\hat{S}(L-K),...\hat{S}(L)\}$$

where $\mathscr{F}^{-1}$ is the inverse of fast Fourier transform, $1 \leq l \leq L$ and K is an even number.

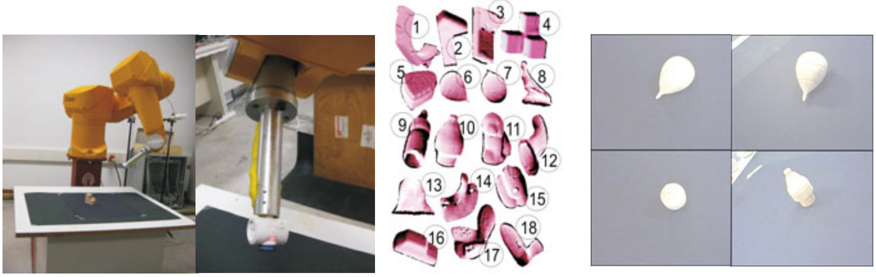
The technical report from [9] provides details of the whole process used to compute the ASSN descriptor and how to estimate the pose parameters between two silhouettes represented using ASSN. The properties required for a shape descriptor: variance, completeness and stabilities, are also proved.

4 Experimentation

In order to prove the effectiveness of the proposed dataset representation model, we have developed a set of experimental tests in a robot-vision system. The experimental setup consisted of a Stabli RX90 robot with a camera on the end-effector of the robot. This vision-robot system was able to capture images around the object placed in the scene. Figure 3(a) illustrates a typical scene with an isolated object placed on a table inside the robot workspace. The tests were carried out on a 3D Synthetic Model library (3DSL)[10]. The experiments were developed by selecting (3DSL) 18 free form objects from this library. This set of objects can be observed in Figure 3(b), while Figure 3(c) shows samples of images captured in the vision-robot system and used during the tests. Note that the object background is uniform but that the objects' illumination changes according to the sensor position.

Shape Recognition Performance of ASSN Descriptor

Since shape recognition is a key process for any 3D recognition system based on object appearance, the first tests have been focused on measuring the performance of our



(a) (b) (c)

Fig. 3. Experimental setup. a) The experimental platform uses a Stabli robot with a camera on the end-effector. b) Synthetic objects in the dataset. c) Samples of images captured in the vision-robot setup and used during the active recognition tests.

suggested shape representation model (ASSN). A comparative analysis was then made, taking the following parameters into consideration: recognition rates ($\mathscr{R}_o$), computational cost ($\mathscr{T}_o$) and pose estimation error ($\mathscr{E}_o$). The pose estimation error is measure as a quadratic mean error between the scene shape and the identified shape in the dataset after to be transformed according the estimated pose parameter. In this test, we compared the behavior of ASSN descriptor with that of other popular descriptors based on shape contours (Fourier descriptor (FD)[11] and Boundary Moment (BM)[12]) and based on shape region (Zernike Moments (ZM)[13] and Complex Moments (CM)[14]. The shape recognition process was developed by using the Euclidean distance as a similarity measure. The shape recognition test is accomplished by capturing one image from a camera located at the end-effector of a robot. We have taken a total of 86 images for this test. Table 1 depicts the average of the parameters under analysis. Upon considering the classical Fourier Descriptor and our suggested ASSN descriptor, the computation cost parameter of ASSN is higher but the recognition rate and pose estimation parameters achieves a better performance. The improvement of ASSN descriptor over traditional Fourier Descriptor is due to the robustness of the ASSN descriptor to shape variations (viewpoint variation, noise, segmentation error). The Complex Moments descriptor is the only descriptor to be compared which has better recognition rates than ASSN. However, it is important to bear in mind that if the object recognition application requires high accuracy for object pose estimation, then the ASSN descriptor is better than the Complex Moment descriptor.

An Active Recognition System Using the Proposed Dataset Representation Model: A Comparative Analysis

To evaluate the performance of the proposed dataset model, we have used the active recognition system developed in [4]. During the comparative process, [4] will be referred as AR 1. Two active recognition systems based on modifications of AR 1 system have been implemented. The first AR 1 modification is based on modeling the objects using [15]. The second AR 1 modification uses the CSS algorithm to model the objects and the object views are represented by ASSN descriptor. Also, we have implemented the active recognition developed by Borosting et.al. [2] based in a probabilistic model and PCA descriptors. From now, AR 1 modified with [15], AR 1 modified with our dataset representation model and Borosting et.al. [2] will be denoted as AR 2, AR 3 and AR 4 respectively. The CSS method reduced the number of views in the dataset to 63% meanwhile, with [15] method the number of views was reduced to 78%.

Table 2 shows the comparison between different shape recognition systems with regard to: recognition rate ($\mathscr{R}_o$) and computational efficiency ($\mathscr{G}_o$). This last parameter is computed by $\mathscr{G}_o = \mathscr{W}_o \cdot \mathscr{T}_o$, where $\mathscr{W}_o$ is the mean sensor positions and $\mathscr{T}_o$ is the computational cost a each sensor position.

From table 2 we can conclude that: (1) The AR 3 system is an improvement of the AR 1 system since the number of sensor positions is reduced, and although the computational cost is increased in one iteration (sensor position) in AR 3, the computational efficiency of this system is higher since it requires a smaller number of sensor positions. This result proves that our representation model is more robust than AR 1 because the shape recognition system identifies a smaller number of hypotheses in each iteration.

Table 1. Comparison between shape recognition systems using different descriptors

Descriptor	$\mathscr{T}_s(s)$	$\mathscr{E}_s(\%)$	$\mathscr{R}_s(\%)$
ASSN	0.18	0.11	66.0
FD	0.14	0.18	58.2
BM	0.09	1.21	23.6
ZM	0.35	2.27	49.9
CM	0.64	2.30	74.1

Table 2. Comparison between different active recognition systems

Active model	$\mathscr{T}_o$ (s)	$\mathscr{W}_o$ $(\%)$	$\mathscr{G}_o$	$\mathscr{R}_o$ $(\%)$
(AR 1)	0.20	5.5	1.10	94
(AR 2)	0.17	5.2	0.88	91
(AR 3)	0.22	3.8	0.85	96
(AR 4)	0.19	4.6	0.90	96

(2) The improvement achieved by AR 2 is lower than by AR 3 since when using the [15] object representation model the number of views in the dataset is reduced in an additional 15%. This behavior is related to the uncertainty that is present during the hypothesis estimation as a result of the differences between the scene view and the key view, and more robot movements are therefore required to solve the uncertainty problem. (3) The performances of AR 4 and AR 3 are very similar, but observe that in the case of AR 3 it is not necessary to use a training step. (4) The comparative study shows the importance of the shape representation model in the computational efficiency parameters since the number of sensor positions is strongly dependent on the robustness in the shape recognition system during the identification of the hypothesis.

In order to evaluate the applicability of the proposed dataset representation model in tasks more complex such as object manipulation, we have used the active recognition AR 3 to develop a simple application: the active recognition system AR 3 must recognize an object in a scene and order the grasping system to pick up the object and leave it in the right box. After 35 tests, the manipulation task was developed in a 97% successfully. The video sample is available at http://isa.esi.uclm.es/descarga-objetos-adan/videoGrasping.wmv.

5 Final Discussions and Conclusions

In this paper we have presented an object dataset representation model for active recognition systems. This model optimizes dataset information. It models the objects by using the proposed Canonical Sphere Section (CSS) model and normalizes the shape to affine transformations and representation accuracy in the spectral dominium (ASSN descriptor). The proposed ASSN descriptor is also robust to small deformations in shape and geometric transformations, and is able to represent, with the same precision, any object shape (silhouette) using the same number of elements in the descriptor. The good performance of the ASSN descriptor has been experimentally proved by comparing its performance with that of other shape descriptors. Our object dataset representation model has also been implemented in an active recognition system and compared with other active recognition systems from the scientific literature, thus showing that it may assist in the construction of more efficient active recognition systems for 3D objects. What is more, the suitability of this representation model in developing object recognition tasks requiring accuracy in the object pose estimation has been tested in an object manipulation robotic application.

Acknowledgement This research was supported by the Spanish Government research Programme via Project DPI2009-09956 (MCyT), by the Junta de Comunidades de Castilla-La Mancha via project PCI-08-0135, and the European Social Fund.

References

1. Kovacic, S., Leonardis, A., Pernus, F.: Planning sequences of views for 3-d object recognition and pose determination. Pattern Recognition 31(10), 1407–1417 (1998)
2. Borotschnig, H., Paletta, L., Pinz, A.: A comparison of probabilistic, possibilistic and evidence theoretic fusion schemes for active object recognition. Computing 62(4), 293–319 (1999)
3. Deinzer, F., Denzler, J., Derichs, C., Niemann, H.: Integrated Viewpoint Fusion and Viewpoint Selection for Optimal Object Recognition. In: British Machine Vision Conference 2006, vol. 1, pp. 287–296 (2006)
4. González, E., Adán, A., Feliú, V., Sánchez, L.: A solution to the next best view problem based on d-spheres for 3d object recognition. In: Proceedings of the Tenth IASTED International Conference on Computer Graphics and Imaging, CGIM 2008, pp. 286–291. ACTA Press, Anaheim (2008)
5. Vranic, D.V.: 3d model retrieval (2004)
6. Tangelder, J.W.H., Veltkamp, R.C.: A survey of content based 3d shape retrieval methods. Multimedia Tools Appl. 39(3), 441–471 (2008)
7. González, E., Adán, A., Feliú, V.: Canonical sphere section: A 3d view-based object representation model. University of Castilla La Mancha, Tech. Rep. (June 2011), http://isa.esi.uclm.es/descarga-objetos-adan/AFFN.pdf
8. Avrithis, Y., Xirouhakis, Y., Kollias, S.: Affine-invariant curve normalization for object shape representation, classification, and retrieval. Machine Vision and Applications 13, 80–94 (2001)
9. González, E., Adán, A., Feliú, V.: Affine spectral silhouette normalized descriptor. University of Castilla La Mancha, Tech. Rep. (June 2011), http://isa.esi.uclm.es/descarga-objetos-adan/AFFN.pdf
10. eii. 3d synthetic library (3dsl) (June 2009), http://eii.unex.es/mallas/
11. González, E., Adán, A., Feliú, V., Sánchez, L.: Active object recognition based on fourier descriptors clustering. Pattern Recogn. Lett. 29, 1060–1071 (2008)
12. Hu, M.-K.: Visual pattern recognition by moment invariants. IRE Transactions on Information Theory 8(2), 179–187 (1962)
13. Khotanzad, A., Hong, Y.: Invariant image recognition by zernike moments. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(5), 489–497 (1990)
14. Flusser, J.: On the independence of rotation moment invariants. Pattern Recognition 33(9), 1405–1410 (2000)
15. Yamauchi, H., Saleem, W., Yoshizawa, S., Karni, Z., Belyaev, A.G., Seidel, H.-P.: Towards stable and salient multi-view representation of 3d shapes. In: SMI, p. 40. IEEE Computer Society, Los Alamitos (2006)

PCA Enhanced Training Data for Adaboost*

Arne Ehlers[1], Florian Baumann[1], Ralf Spindler[2],
Birgit Glasmacher[2], and Bodo Rosenhahn[1]

[1] Institut für Informationsverarbeitung,
Leibniz Universitüt Hannover,
{ehlers,baumann,rosenhahn}@tnt.uni-hannover.de
[2] Institut für Mehrphasenprozesse,
Leibniz Universität Hannover

Abstract. In this paper we propose to enhance the training data of boosting-based object detection frameworks by the use of principal component analysis (PCA). The quality of boosted classifiers highly depends on the image databases exploited in training. We observed that negative training images projected into the objects PCA space are often far away from the object class. This broad boundary between the object classes in training can yield to a high classification error of the boosted classifier in the testing phase. We show that transforming the negative training database close to the positive object class can increase the detection performance. In experiments on face detection and the analysis of microscopic cell images, our method decreases the amount of false positives while maintaining a high detection rate. We implemented our approach in a Viola & Jones object detection framework using AdaBoost to combine Haar-like features. But as a preprocessing step our method can easily be integrated in all boosting-based frameworks without additional overhead.

1 Introduction

Several well known algorithms exist to perform object detection and recognition in images. For an overview on existing approaches and databases in the field of face detection, we recommend the web page [7] or the overview articles [17] and [18]. In the vast amount of available techniques, two complementary strategies are very common for object detection, namely boosting [6,13] and PCA-based methods [12,4]. Simply speaking and more detailed in Section 2.1, the strategy behind AdaBoost is to linearly combine several weak classifiers to gain a strong classifier. By using integral images, the classifiers are based on the difference of local (rectangle shaped) image patches. Object detection using AdaBoost is known to be slow in the training phase, but real-time capable in detection, even on resource limited systems. AdaBoost is further known to be sensitive in generating false-positive mistakes, which is sometimes compensated for by using

* This work has been partially funded by the DFG within the excellence cluster REBIRTH.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 410–419, 2011.

additional post-processing steps, such as Canny-pruning or histogram analysis. In contrast to a linear combination of local classifiers, the idea behind a PCA-space of objects is to learn a global subspace from training data which span the object variations as principal components in a high-dimensional vector space [12]. Additional to the PCA, several variants, e.g. independent component analysis (ICA) [2] or Kernel PCA (KPCA)[11] have also been proposed for face detection and recognition. The method for (K)PCA-based object detection is explained in more detail in Section 2.2. PCA based methods are more robust to noise since a subspace is learned from the training data, but much slower in the detection phase, because image patches need to be projected in the object space. For this reason we decided not to boost features evaluated in PCA-space as proposed in [19] and [1].

Since boosting- and PCA-based approaches rely on completely different principles, our main interest is to combine both methods without introducing additional limitations. So the key question is how to integrate a learned subspace of objects (e.g. faces, cells) in the boosting approach. Our observation is, that training data for e.g. non-objects are commonly very far away from the object space, once a PCA-space has been learned from the training data.

So the key contribution described in Section 3 is to modify the training-data of the non-objects in such a way, that they are closer to the PCA-space of objects and therefore to cause a much smaller margin at the start of training. This is in some way contrary to approaches in semi-supervised learning [8] in which the negative training class is raised at the boundary being opposite to the object class.

Figure 1 shows two non-object examples which are morphed towards their object-space, namely a non-face towards face-space in the top row and a defective cell towards cell-space in the bottom row. The middle images are non-objects which are much better suited to learn a boundary between the positive and negative classes in the boosting framework. Overall, it allows to train a much more selective classifier, especially if only a sparse amount of training data is available. Additionally, since the amount of training data remains unchanged (images of non-objects are replaced with synthesized new images closer to the object space), we do not introduce any additional overhead in the training or testing phase in

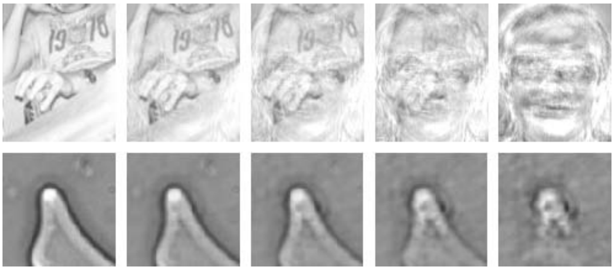


Fig. 1. Morphing a non-object towards the a PCA-trained object space. Top : Example morph for a face space. Bottom: Example morph for a cell space.

conventional AdaBoost. Another advantage is, that many existing modifications and improvements to the classical AdaBoost algorithm such as Entropy Regularized LPBoost, SoftBoost, MILBoost or SEAdaboost [16,15,14,3] can still be used, since we only perform a kind of pre-processing with the training data (in this case for the non-objects). Therefore, it is sufficient to use the conventional AdaBoost algorithm to evaluate the impact of our approach.

2 Foundations

This section introduces the foundations for our work, namely the AdaBoost algorithm and the basic idea behind PCA-learning.

2.1 AdaBoost

AdaBoost is a popular machine learning algorithm proposed by Freund and Schapire [6] that can be applied to achieve good results in different areas in object detection. As a boosting algorithm it forms a strong classifier while trained on labeled classes of training data. The boosted classifier is a linear combination of single classifiers selected from a given set in each training round due to their minimal classification error. This error is calculated depending on weights assigned to the training images. By adapting these weights AdaBoost concentrates in later training rounds on the examples that are hard to classify. Suitable as weak classifiers are Haar-like features because of their fast and simple computation as proposed by Viola and Jones in [13]. Because of AdaBoost's autonomous learning the quality of the boosted classifier depends to a high degree on the characteristics of the training classes.

2.2 PCA-Learning

A principal component analysis (PCA) transforms (possibly) correlated variables into uncorrelated variables called principal components by applying an orthogonal transformation. The transformation is designed in such a fashion that the first axis (principal component) represents the highest amount of data variation, whereas the other following (orthogonal) axes are sorted in a decreasing order, depending on the amount of variance. A PCA is simply computed as a singular value decomposition of a data matrix or by an eigendecomposition of a covariance matrix generated from a data set. In our implementation we derived the eigenvectors from a singular value decomposition of the covariance matrix. As that covariance matrix is symmetric and positive semidefinite the obtained eigenvectors are identical to those provided by a PCA and also the order of the corresponding eigenvalues is the same.

Following the notations in [12], we assume n-dimensional data points $\Gamma_1 \dots \Gamma_m$ and compute the average of the data points as

$$\Psi = \frac{1}{m} \sum_{i=1}^{m} \Gamma_i \tag{1}$$

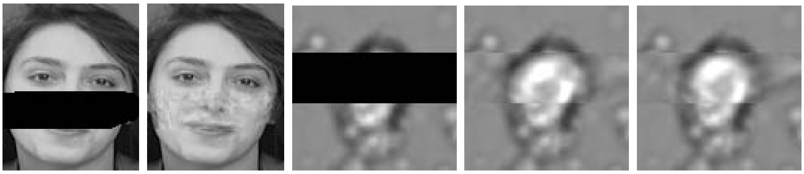


Fig. 2. Missing data estimation of face image (on the left) and a cell image (on the right). No blending was performed and it is shown, that the mouse and nose as well as the cell structure is well approximated. The reconstruction of the cell to the right is based on a Kernel-PCA. For this kind of data, it seems to approximate the cell boundaries slightly better.

We further define $\Phi_i = \Gamma_i - \Psi$ as the difference to the average vector (and shift the data in this way towards the origin). The matrix $A = [\Phi_1, \ldots, \Phi_m]$ contains the difference vectors of the data, so that the covariance matrix of the data points is given as

$$C = \frac{1}{m} \sum_{i=1}^{m} \Phi_i \Phi_i^T \tag{2}$$

$$= AA^T \tag{3}$$

A singular value decomposition $C = UDV^T$ of the covariance matrix C allows to compute the principal components of the data. Note, that the computation of $A^T A$ can be much more efficient, if less data points are available than the dimension of the data (see [12] for details). This happens, e.g. when face images are encoded as vector and only a few face images (e.g. a couple hundred) are available. Then U needs to be multiplied by A and rescaled to get the eigenvectors. Then the first s Eigenvectors can be used for approximation of the face space. Figure 2 shows two simple examples in which the missing parts of a corrupted face and cell image (not contained in the training data) has been reconstructed with the help of a database by performing a subspace projection. As can be seen, the position of the nose and mouth as well as the cell boundaries have been reconstructed fairly well.

There exist many extensions and modifications about PCA-methods for data clustering. In our experiments on Cell-data we will use a Kernel-PCA (KPCA) [11]. Here, we want to demonstrate that the method on PCA-enhancement of training data is not restricted to a specific method for subspace learning. The idea for KPCA is to employ a mapping $\phi(x)$ on the data to lift the input to a higher dimensional space, in which the subspace can be approximated more easily. Since the higher dimensional space can become very large, they key idea behind Kernel-PCA is to avoid the explicit computation of ϕ and to work with a kernel $K_{i,j} = k(x_i, x_j) =< \phi(x_i), \phi(x_j) >= \phi(x_i)^T \phi(x_j)$. The covariance matrix C then becomes

$$C = \frac{1}{m} \sum_{i=1}^{m} \phi(x_i) \phi(x_i)^T \tag{4}$$

which is again splitted and normalized after a SVD, $C = UDV^T$. In our experiments we used two standard kernels, namely $K(x, y) = -\exp((x - y)^2/(\sigma)^2)$ with $\sigma = 1$ and $K(x, y) = (x^T y)^2$.

3 Training Data Enhancement

Obviously one possibility to combine PCA-based face detection and boosting is to use both algorithms for classification separately and to join (in a smart way) the outcome. The disadvantage of this method would be that the superior time performance of the boosting approach would be lost since the computation of the PCA-mapping takes significantly longer time for a test image. Therefore, the idea is to modify and enhance the training data. The impact of some characteristics of the training data on the test error is subject to the margin theory, that is described briefly in the following.

3.1 Margin Analysis

Shortly after publication of the AdaBoost algorithm research has been started to examine its detection performance. As the consideration of only the training error is not sufficient to estimate the test error of AdaBoost, Schapire et al. [10] proposed the margin as a measure of the confidence in the algorithms classification. They defined the margin as the difference between the sum of the weights of the weak classifiers voting for the correct object class and the maximal sum of weights assigned to an incorrect class. As these weights are normalized to sum up to one, the margin is defined in the range [-1,1] and a positive value implies a correct decision. Hence a large positive margin represents a confident correct classification.

Evaluated on the complete training set a margin distribution can be derived. Schapire et al. observed that, due to its adaption of the training example weights, AdaBoost proceeds very aggressive in reducing the amount of training examples having a small margin. For this reason the AdaBoost learning algorithm can reduce the test error even after the training error has reached zero.

In the last decade much research has been done in estimating the test error subject to the margin and find a boundary based on the minimum margin and other training parameter. But recent research [9] indicates as well that considering the minimum margin is not sufficient and also the complete margin distribution has to be taken into account.

The margin theory should not be discussed in this work in detail. As we followed a more empirical approach in enhancing our detection framework, an experiment on the impact of our method on the margin distribution will be presented in Section 4.1.

3.2 Training Data Adaption

In our experience, negative training examples are not always well chosen to differ between objects and non-objects. This is mainly due to the fact, that the non-object space is significantly larger and more complex than the positive examples.

Basically, the non-objects can be seen as the complementary space to the learned PCA object space. Therefore, its variability is hard to reflect in the training data. The idea is to bring the training data close to the PCA-space. This can simply be done by projecting the negative training examples onto the trained PCA-space and then shifting it back towards the non-object space with a scale $\lambda \in [0 \dots 1]$: Let $U_s = U(1 : n, 1 : s), s \leq n$ be the upper left matrix of U stemming from $C = UDV^T$ and let T be an example of the non-object class. The shift of T towards T_s being closer to the object space can simply be done by computing

$$Proj = U_s^T \cdot (T - \Psi) \tag{5}$$

$$Rec = U_s \cdot Proj + \Psi \tag{6}$$

$$T_s = Rec + \lambda(T - Rec) \tag{7}$$

Note, that U is an unitary matrix and thus U_s describes the inverse projection of U_s^T. In case of the more efficient computation mentioned in Section 2.2 this property is not given and instead a pseudoinverse has to be used. Obviously, $\lambda = 0$ yields the projection on the PCA-space, whereas $\lambda = 1$ leads to the training example itself. So λ steers the amount on how much the example is shifted towards the object space. In our experiments we used $\lambda = 0.3, 0.5$ and 0.7, respectively. Some example morphs of negative training data for different weighting factors are shown in Figure 1 and the second row of Figure 3.

4 Experiments

We decided for two kinds of experiments. The first experiments are on face detection. Here we use the well known AT&T database of faces. Therefore we give credits to AT&T Laboratories Cambridge. The database contains ten different images each of 40 people varying in the lighting, facial expressions and facial details (glasses / no glasses). The images were taken against a dark homogeneous background with the people in an upright, frontal position.

The second set of experiments is performed on microscopic images of cells. The cells are recorded during cryo-conservation, and therefore ice fronts are forming around the cells. The goal is to detect (and track) the cells in the videos. Here we collected 250 images of cells and 350 images of non-cells, so that we gain a reasonable database. We divided our image bases into a training and validation set using a 67/33 ratio. Crowther and Cox [5] illustrated that especially for small bases a split containing only a small part for validation is not recommendable. They suggested to select a ratio between 50/50 and 70/30.

Figure 3 shows in the top row example images of non-faces and non-cells. The middle row shows morphed images towards the trained PCA space. These images are then used for training to find a more selective classifier. The bottom row shows positive example images of faces and cells of the used databases.

4.1 Experiments with the AT&T Database

Using the PCA from Section 3 we generated four different training sets containing the good/bad examples and morphed bad examples with $\lambda = 0.3, 0.5$ and 0.7.

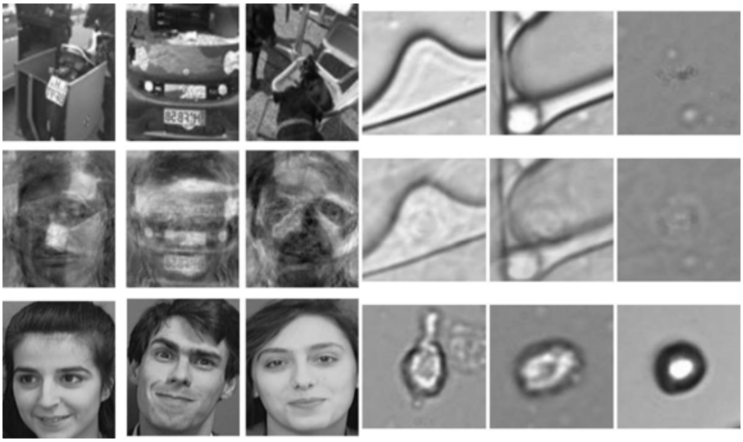


Fig. 3. Top row: Example images of non-faces and non-cells. Middle row: Morphed images towards the trained PCA space. These images are either used to find a more selective classifier. Bottom row: Example faces and cells of the used databases.

For all four data sets we performed an AdaBoost learning as described in Section 2.1. Here we used simple rectangular features for training.

Margin distribution. Figure 4(a) presents the cumulative margin distributions of the original face training set after different training rounds. In Figure 4(b) the margins boosting the AT&T database using PCA-enhanced non-faces with $\lambda = 0.3$ is shown. The impact of the boosting process on the margin distribution is clearly noticeable in both figures. The amount of training examples having a small margin is in both cases strongly reduced during training. Roughly after 10 rounds the training error reaches zero as all examples images have a positive margin and hence are correctly classified. Then the AdaBoost algorithm further concentrates on the training examples that are hard to classify and continues to reduce the number of narrow decisions.

But it is also observable that it is more difficult to classify the morphed training set. After 5 training rounds the boosted classifier for the morphed set makes almost twice as much wrong decisions compared to the classifier boosted on the original training set. Also about 15% and 30% of the training examples on the morphed set have a margin smaller than 0.2 and 0.56, respectively. In comparison, for the original set the smallest 15% have a margin below 0.32 and the margins of the smallest 30% do not exceed 0.62.

After 40 training rounds the boosted classifier for the morphed training set has caught up in the lower region of the margin distribution. The minimum margin amounts roughly to 0.26 in both cases and the progress of cumulative distributions is similar showing only a slightly steeper slope for the morphed training set.

As discussed in Section 3.1 the margin distribution in training has been found to be an indicator for the quality of a classifier in terms of its test error. Hence

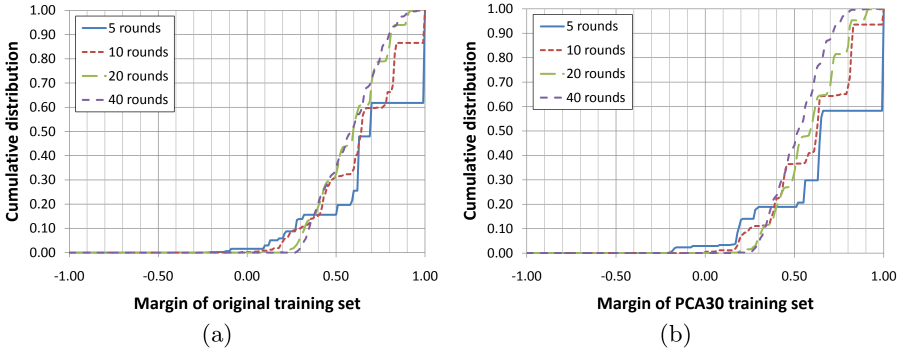


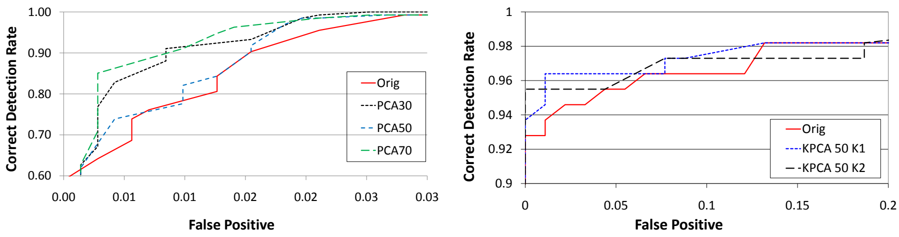
Fig. 4. (a): Cumulative margin distributions of the original face training set after 5, 10, 20 and 40 rounds. (b): Cumulative margin distributions of the morphed training set. The negative object class has been morphed using $\lambda = 0.3$.

the result of the PCA enhanced training to achieve a similar margin distribution starting from an adverse one can be interpreted as a higher gain during training. Therefore we expect classifiers boosted using PCA enhanced training data to achieve superior performance in the test phase.

Face detection. In the following the results of experiments on the test set are presented. Figure 5(a) shows the ROC-curves of multiple classifiers varying the classification threshold in detecting faces. The curve in red is the classifier based on the original data set, whereas the other curves show the performance of the classifiers using PCA-images for training. Overall, the curves show that the classifiers which have been trained with the PCA-images are more selective in detecting faces so that good detection rates are achieved while maintaining a lower false alarm rate.

4.2 Experiments with the Cell Database

For the cell database we decided on using a Kernel-PCA-method for modifying the training data. The main reason is to demonstrate that variants of PCA-learning can be used in a similar fashion. Especially for image data with larger image gradients (due to edges), KPCA-methods can be better suited, since the overall smoothing effect using PCA can be reduced. Since the KPCA-enhanced training is dependent on the selection of the kernel function we further decided to compare two different (standard) kernels, namely $K_1(x,y) = -\exp((x-y)^2/(\sigma)^2)$ and $K_2(x,y) = (x^T y)^2$. The KPCA-enhanced training data leads for both kernels to an increased performance of the detection rate, which is shown in the ROC-curve in Figure 5(b). E.g. for a detection rate of 96.3%, the PCA-enhanced training data with K_1 yields a classifier which produces a false-positive detection rate of 1%, whereas the original data produces a false-positive detection



(a) (b)

Fig. 5. (a): ROC curve for the face database using different thresholds of boosting with the original data (red) and using PCA-enhanced non-faces with different λ-values (0.3, 0.5 and 0.7). The PCA-enhanced data reveals a much more selective performance. (b): ROC-curve for the cell database.

rate of 6%. The PCA-enhanced training data with K_2 yields similar performance, in producing a detection rate of 95.5% with no false positives.

5 Conclusion

We introduced an approach to enhance the training data of boosting-based object detection frameworks to achieve a higher detection performance. Using principal component analysis we shift the negative training examples in PCA space near to the positive training class. The trained classifier achieves a lower classification error being more selective in detection. Our experiments on face detection and microscopic cell images showed that our method decreases the false positive rate of the boosted classifier. The variable strength of our transformation allows for a trade-off between true positive and false alarm rates. But in all experiments our approach managed to significantly lower the amount of false alarms without reducing the detection rate. As a preprocessing of the training data our method can be integrated in nearly all boosted detection frameworks without any computational costs in the learning and detection process.

References

1. Ali, S., Shah, M.: An integrated approach for generic object detection using kernel pca and boosting. In: ICME, pp. 1030–1033 (2005)
2. Bartlett, M., Movellan, J., Sejnowski, T.: Face recognition by independent component analysis. IEEE Transactions on Neural Networks 13(6), 1450–1464 (2002)
3. Baumann, F., Ernst, K., Ehlers, A., Rosenhahn, B.: Symmetry enhanced adaboost. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Chung, R., Hammoud, R., Hussain, M., Kar-Han, T., Crawfis, R., Thalmann, D., Kao, D., Avila, L. (eds.) ISVC 2010. LNCS, vol. 6453, pp. 286–295. Springer, Heidelberg (2010)
4. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: SIGGRAPH 1999: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, pp. 187–194. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA (1999)

5. Crowther, P.S., Cox, R.J.: A method for optimal division of data sets for use in neural networks. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3684, pp. 1–7. Springer, Heidelberg (2005)
6. Freund, Y., Schapire, R.E.: A short introduction to boosting. Journal of Japanese Society for Artificial Intelligence 14(5), 771–780 (1999)
7. Homepage, F.D.: (2010), http://www.facedetection.com/
8. Leistner, C., Grabner, H., Bischof, H.: Semi-supervised boosting using visual similarity learning. In: CVPR (2008)
9. Li, H., Shen, C.: Boosting the minimum margin: Lpboost vs. adaboost. In: Proceedings of the International Conference on Digital Image Computing: Techniques and Applications, DICTA, pp. 533–539 (2008)
10. Schapire, R.E., Freund, Y., Barlett, P., Lee, W.S.: Boosting the margin: A new explanation for the effectiveness of voting methods. In: Proceedings of the Fourteenth International Conference on Machine Learning (ICML), pp. 322–330 (1997)
11. Schölkopf, B., Mika, S., Smola, A., Rätsch, G., Müller, K.R.: Kernel pca pattern reconstruction via approximate pre-images. In: Proceedings of the 8th International Conference on Artificial Neural Networks, Perspectives in Neural Computing, pp. 147–152. Springer, Heidelberg (1998)
12. Turk, M., Pentland, A.: Face recognition using eigenfaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–591. IEEE Computer Society, Los Alamitos (1991)
13. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004)
14. Viola, P., Platt, J.C., Zhang, C.: Multiple instance boosting for object detection. Advances in Neural Information Processing 18, 1417–1426 (2007)
15. Warmuth, M.K., Glocer, K., Raetsch, G.: Boosting algorithms for maximizing the soft margin. Advances in Neural Information Processing Systems 20, 1585–1592 (2008)
16. Warmuth, M.K., Glocer, K.A., Vishwanathan, S.: Entropy regularized lpboost. In: Freund, Y., Györfi, L., Turán, G., Zeugmann, T. (eds.) ALT 2008. LNCS (LNAI), vol. 5254, pp. 256–271. Springer, Heidelberg (2008)
17. Yang, M.H., Kriegman, D.J., Ahuja, N.: Detecting faces in images: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 34–58 (2002)
18. Zhang, C., Zhang, Z.: A survey of recent advances in face detection. Microsoft Research Technical Report, MSR-TR-2010-66 (2010)
19. Zhang, D., Li, S.Z., Gatica-Perez, D.: Real-time face detection using boosting in hierarchical feature spaces. In: ICPR, vol. (2), pp. 411–414 (2004)

Psychophysically Inspired Bayesian Occlusion Model to Recognize Occluded Faces

Ibrahim Venkat[1], Ahamad Tajudin Khader[1],
K.G. Subramanian[1], and Philippe De Wilde[2]

[1] School of Computer Sciences, Universiti Sains Malaysia, Malaysia
ibrahim@cs.usm.my
[2] School of Math. & Comp. Sciences, Heriot-Watt University, Edinburgh, UK

Abstract. Face recognition systems robust to major occlusions have wide applications ranging from consumer products with biometric features to surveillance and law enforcement applications. In unconstrained scenarios, faces are often subject to occlusions, apart from common variations such as pose, illumination, scale, orientation and so on. In this paper we propose a novel Bayesian oriented occlusion model inspired by psychophysical mechanisms to recognize faces prone to occlusions amidst other common variations. We have discovered and modeled similarity maps that exist in facial domains by means of Bayesian Networks. The proposed model is capable of efficiently learning and exploiting these maps from the facial domain. Hence it can tackle the occlusion uncertainty reasonably well. Improved recognition rates over state of the art techniques have been observed.

Keywords: Face Recognition, Occlusion Models, Similarity Measures, Bayesian Networks, Parameter Estimation.

1 Introduction

Substantial research is ongoing in both the fields viz., cognitive psychology and machine vision to understand the effect of occlusions in recognizing human faces. Till date the problem of recognizing faces prone to occlusion appears to be a partially solved problem in computer vision [1,2,3]. In this paper we propose a novel *"**P**sychophysically **I**nspired **B**ayesian **O**cclusion **M**odel* (**PIBOM**)" to attack this potential problem.

The formulation of PIBOM has been inspired by some key cognitive psychology principles which we briefly describe here. Similarity is a basic concept in cognitive psychology which is utilized to explore the principles of human perception [4]. Recent studies [5,6] refer to the classical contrast model of similarity [7] which insists that perceived similarity is the result of a feature-matching process. A fundamental hypothesis [8] associated with the perception and memory of faces states that, humans perceive and remember faces chiefly by means of facial features. Psychological experiments [8] that evaluate similarity judgments support this hypothesis. Facial processing algorithms used by

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 420–426, 2011.

popular imaging applications such as *photofit* and *identikit* are based on this cognitive phenomenon. PIBOM as well, is based on these fundamental insights about human pattern matching and memory. While reasoning with objects which are prone to uncertainties, humans are often able to notice similarities between subregions of a face and a set of faces. For example when a face is cluttered with occlusions, we would be able to recall some individuals by just observing a particular portion of the face which closely resembles the characteristic of those individuals. PIBOM precisely intends to map intrinsic similarities between the set of subsamples of a given probe face with the set of faces in the facial domain by means of Bayesian Networks(BNs). We briefly present the three-fold architecture of our PIBOM with the aid of the flow diagram shown in Fig. 1. Firstly, we normalize the face images using the technique proposed by Bartlett et al.[9]. Then the feature space (low dimensional eigenspace) is constructed from the gallery (training set) of face images available in the *Face DataBase* (FDB). PIBOM further learns the belief states of various facial subregions from the enhanced FDB using a standard machine learning procedure called *Parameter Estimation.* All these preliminary activities in *Phase I* have been done off-line to make minimal use of computing resources. In *Phase II* the probe face is similarly enhanced and subject to horizontal segmentation. Then the PCA features of facial entities, acquired by combining probe face components over the gallery face images, are extracted and projected over the feature space using an inheritance mechanism which will be described in section 3. Further, probable subjects are shortlisted by means of comparing similarity measures. In *Phase III*, a BN is generated for a given probe, whose child node variables represent the belief states of short-listed subjects and the parent nodes represent the belief states of

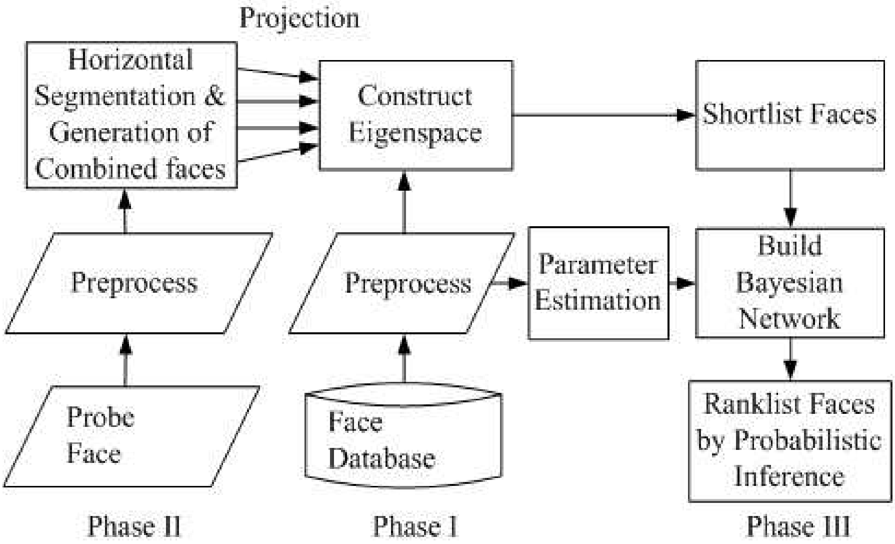


Fig. 1. Flow diagram of the proposed PIBOM

corresponding components of the probe face. Finally faces are rank-listed using a face score function which will be defined in section 3.

2 Related Work

Psychologically feasible computational models exhibit clear and strong relationships between behavior and properties of the domains which they intend to represent[10]. The psychophysical experiments performed by Schwaninger et al. [11] supports the notion that component based processing of faces is preferable than global processing, which is orientation sensitive unlike component based processing. Behrmann and Mozer [12] have shown that humans, in order to minimize the processing load, organize a complex occluded object into subregions and then attend selectively to particular physical regions. A very recent psychological study [13] empirically shows that facial identity information is conveyed largely via mechanisms tuned to horizontal visual structure. Based on these psychological evidences, the proposed PIBOM strategically uses horizontal processing.

The Martinez Localization Algorithm (MLA) [1], which serves as a baseline occlusion model, attempts to recognize partially occluded faces with frontal views using a PCA based approach. It demands high computation time due to the use of mixtures of Gaussian distributions. This approach mainly tackles frontal view images without any pose variations. Like MLA, the proposed PIBOM does not impose a restriction that the face images should be frontal. Bayesian models have been applied in a variety of applications that work in unconstrained and realistic environments and they are now the mainstay of the AI research field known as "uncertain reasoning" [14].

3 Formulation of PIBOM from the Facial Domain

Let the probe face be segmented into k equal horizontal rectangular subregions. Let $S = \{S_1, S_2, S_3, \cdots, S_k\}$ represent the k subsamples of the probe face. Let $F = \{F_1, F_2, F_3, \cdots, F_n\}$ represent the training face set which has face images of n subjects. Let us suppose that, a typical subsample $S_i \subset S$, $1 \leq i \leq k$, might influence the recognition of a set of faces $f = \{F_p, F_q, F_r\} \subset F$, where p, q and r represent unique integers between 1 and n. Let Z_{ip}, Z_{iq} and Z_{ir} represent the corresponding influence strengths. Since S_i is influencing the recognition of f, we draw edges from S_i to the elements of f, resulting in a typical Directed Acyclic Graph (DAG) as shown in Fig.2. By this way we establish mappings from the set of subsamples (S_i) to the subset of faces activated (f). Conceptually these faces will be nearly similar to the probe face which represents these subsamples. In the DAG shown in Fig.2 each face is conditionally independent of the other faces given its parent. That is $I_P(\{F_p\}, \{F_q, F_r\}|S_i), I_p(\{F_q\}, \{F_r, F_p\}|S_i)$ and $I_p(\{F_r\}, \{F_p, F_q\}|S_i)$, where we denote independence of random variables by I_P. This can be be precisely written in the following general form

$$P(F_j|F_c, S_i) = P(F_j|S_i), \quad i = 1, \ldots, k, \quad j = 1, \ldots, n. \tag{1}$$

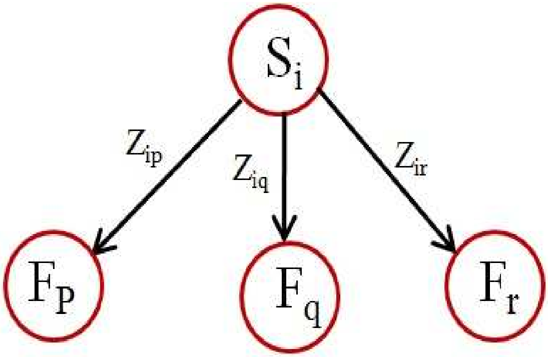


Fig. 2. Proposed PIBOM showing mappings between a subsample and the faces being recognized as a consequence of its influence

where $F_c = F \setminus F_j$. The graphical nature of PIBOM can help us to visualize the abstract intrinsic similarity relationships that exists in a facial domain, as a consequence of mapping S_i to f.

Inheriting similarity mappings from the subspace: We intend to shortlist r similar faces (closely resembling the probe face) from the FDB which is a consequence of the influence of the subregions of the probe face. This will aid us to predict the faces influenced by the horizontal subregions of the probe face by inheriting the PCA architecture. Please note that the proposed PIBOM can be fitted into any suitable subspace projection technique (eg. PCA, ICA, LDA and so on). As an example we have chosen the well known PCA architecture. As the eigenspace is built with the eigenfaces, we cannot directly project the subsamples S_i which do not represent the whole face into this subspace. We strategically combine the subsamples into each of the faces in the *Face Database (FDB)* and project this combined face, say X_{ij}, onto the eigenspace, where X_{ij} is given by

$$X_{ij} = S_i \cup F_j, \quad i = 1, \dots, k, \quad j = 1, \dots, n. \tag{2}$$

Let $SM(F_i, F_j) \in [0, 1]$ represent the similarity measure between two faces F_i, F_j. Then, faces influenced (FI) by subsamples can be computed by

$$FI = \arg\min_{F_j} SM(X_{ij}, F_j), \qquad i = 1, \dots, k, \qquad j = 1, \dots, n. \tag{3}$$

We can project the combined face X_{ij} into the eigenspace using,

$$\omega_{ij} = u_j^T (X_{ij} - \Psi), \qquad i = 1, \dots, k, \qquad j = 1, \dots, n, \tag{4}$$

where ω_{ij} , u_j and Ψ are respectively the weight vectors, eigenvectors and the mean face of the FDB. The face space projection Φ_f can be computed by

$$\Phi_f = \sum_{j=1}^{n} \omega_{ij} u_j, \qquad i = 1, \dots, k. \tag{5}$$

The Euclidean distance between X_{ij} and the face space projection can be computed using

$$\epsilon_{ij} = \| (X_{ij} - \Psi) - \Phi_f \| \tag{6}$$

Let E_s represent the sorted Euclidian distances of ϵ_{ij}. Consequently the r face classes that correspond to the first r Euclidean distances of E_s will yield the faces influenced by each of the horizontal subregions of the probe face. Similar to how a human recalls some faces by observing portions of a face, the above formulation aids the machine to shortlist faces by observing subsamples of a face via psychophysical means. We mathematically define the influence strength Z_{ij} of a subsample S_i as

$$Z_{ij} = (n - \ell)/n \tag{7}$$

where ℓ is the rank in which the face F_j is being recognized by the subsample S_i. We perform the standard *Maximum Likelihood Estimation (MLE)* for all subsamples $i = 1, \ldots, k$, to estimate the parameters Z_{ij} that best agrees with the observed gallery set of face images. Using a gradient method, a set of necessary conditions for the maximum-likelihood estimate for Z_{ij} can be obtained from the set of k equations

$$\sum_{j=1}^{n} \nabla_{Z_i} \ln P(F_j|S_i) = 0 \qquad i = 1, \ldots, k \tag{8}$$

where the gradient operator ∇_{Z_i} is given by

$$\nabla_{Z_i} \equiv \begin{pmatrix} \frac{\partial}{\partial Z_{i1}} \\ \frac{\partial}{\partial Z_{i2}} \\ \vdots \\ \frac{\partial}{\partial Z_{in}} \end{pmatrix} \qquad i = 1, \ldots, k \tag{9}$$

By utilizing the crucial influence strengths to weigh the prior probability of subsamples, $P(S_i)$, we can define the face score μ of a m^{th} face conditioned on S_i using

$$\mu(F_m) = \sum_{S_i} P(S_i)P(F_m|S_i) + \sum_{S_i} Z_{im}P(S_i) \tag{10}$$

This face score function has been used to rank-list the probable faces influenced by the subsamples of a given probe face.

4 Experimental Results and Discussions

We have implemented PIBOM using the MATLAB based *Bayesian Net Toolbox*, developed by Dr.Murphy's team, University of British Columbia. We have used the huge AR FDB [15] which consists of over 3200 face images for our experiments. The dataset has face images with varying facial expressions, illumination conditions and occlusions. Also it offers duplicate probe images, which were taken after a gap of 14 days. Though we use four face images for training, we treat all the training samples as different classes and during testing we consolidate the resulting ranks of recognized face images based on their first occurrence. For example, let the training samples of class i be denoted as $Ci1, Ci2, Ci3$ & $Ci4$ and say, $C13, C11, C22, C14, C12$ & $C32$ have been rank-listed. Then the consolidated ranking will be $C1, C2$ & $C3$. If the actual gallery match (true class) is $C3$ then rank-3 classifier will register a match. This strategy would enable the proposed PIBOM to be compared with models such as MLA that attempt to use single training sample per class. The comparative results of PIBOM with MLA against two typical real occlusions viz., sunglass and scarf, in terms of Cumulative Match Characteristics are as shown in Figs. 3, 4. For the first few ranks MLA leads. However, we see that PIBOM eventually outperforms MLA. It is reported in [1] that recognition tests on non duplicate images are tougher. Even against these tougher tests, PIBOM reports promising recognition rates of about 95% within 4 to 11 ranks, considering the overall performance of all the experiments.

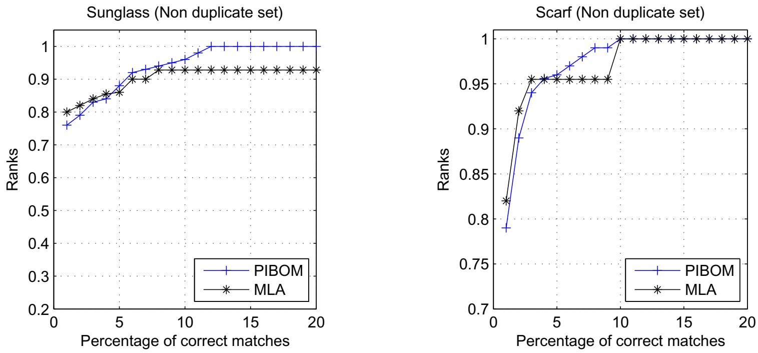


Fig. 3. Comparative analysis of PIBOM, and MLA using the non-duplicate AR dataset

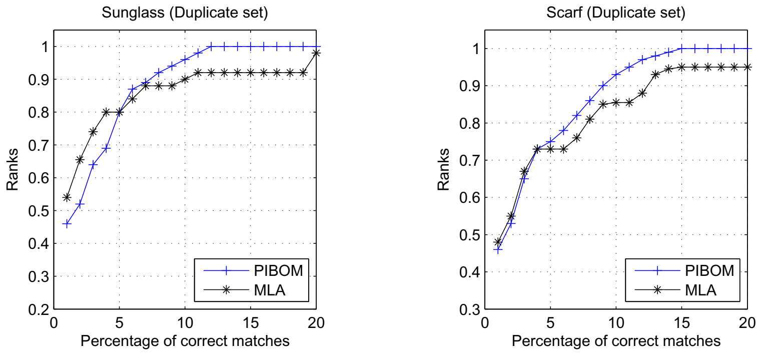


Fig. 4. Comparative analysis of PIBOM and MLA using the duplicate AR dataset

5 Conclusion and Future Work

We have discovered that faces exhibit interesting similarity mappings and successfully modeled an intuitive Bayesian approach to tackle the occlusion problem. In the near future we intend to extend the proposed PIBOM over a wider range of object recognition problems where some uncertainty issues throw major challenges.

Acknowledgments. We acknowledge the anonymous reviewers for providing valuable suggestions. Part of the work has been supported by the incentive grant offered by University Science Malaysia. We duly acknowledge School of Mathematical and Computer Sciences (MACS) , Heriot-Watt University for extending their collaborative support.

References

1. Martinez, A.M.: Recognizing imprecisely localized, partially occluded and expression variant faces from a single sample per class. IEEE Trans. Pattern Analysis and Machine Intelligence 24(6), 748–763 (2002)
2. Kim, J., Choi, J., Yi, J., Turk, M.: Effective representation using ICA for face recognition robust to local distortion and partial occlusion. IEEE Trans. Pattern Analysis and Machine Intelligence 27(12), 1977–1981 (2005)
3. Kanan, H.R., Faez, K.: Recognizing faces using Adaptively Weighted Sub-Gabor Array from a single sample image per enrolled subject. Image & Vision Computing 28(3), 438–448 (2010)
4. Solan, Z., Ruppin, E.: Similarity in perception: A window to brain organization. Journal of Cognitive Neuroscience 13, 18–30 (1999)
5. Nosofsky, R.M., Zaki, S.R.: A hybrid-similarity exemplar model for predicting distinctiveness effects in perceptual oldnew recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition 29, 1194–1209 (2003)
6. Solar, J.R., Navarrete, P.: Eigenspace-based Face Recognition. Springer Engineering Series (2001)
7. Tversky, A.: Features of similarity. Psychological Review 84, 327–352 (1977)
8. Rakover, S.S.: Featural vs. configurational information in faces: A conceptual and empirical analysis. British Journal of Psychology 93, 1–30 (2002)
9. Bartlett, M.S., Movellan, J.R.: Face recognition by independent component analysis. IEEE Trans. on Neural Networks 13(6), 1450–1464 (2002)
10. Fific, A.: Emerging holistic properties at face value: assessing characteristics of face perception, Ph.D. thesis (2005)
11. Schwaninger, A., Wallraven, C., Bulthoff, H.H.: Computational modeling of face recognition based on psychophysical experiments. Swiss Journal of Psychology 63(3), 207–215 (2004)
12. Behrmann, M., Zemel, R.S., Mozer, M.C.: Object-based attention and occlusion. Journal of Experimental Psychology: Human Perception and Performance 24(4), 1011–1036 (1998)
13. Dakin, S.C., Watt, R.J.: Biological bar codes in human faces. Journal of Vision 9(4), 1–10 (2009)
14. Jensen, F.V., Nielsen, T.D.: Bayesian Networks and Decision Graphs. Springer, Heidelberg (2007)
15. Martinez, A.M., Benavente, R.: The AR Face Database. CVC technical report no. 24, Tech. rep. (1998)

Unsupervised Feature Selection and Category Formation for Generic Object Recognition

Hirokazu Madokoro, Masahiro Tsukada, and Kazuhito Sato

Department of Machine Intelligence and Systems Engineering,
Akita Prefectural University,
84–4 Aza Ebinokuchi Tsuchiya, Yurihonjo City, Akita, Japan
madokoro@akita-pu.ac.jp

Abstract. This paper presents an unsupervised method for selection of feature points and object category formation without previous setting of the number of categories. For unsupervised object category formation, this method has the following features: selection of target feature points using One Class-SVMs (OC-SVMs), generation of visual words using Self-Organizing Maps (SOMs), formation of labels using Adaptive Resonance Theory-2 (ART-2), and creation and classification of categories for visualizing spatial relations between them using Counter Propagation Networks (CPNs) . Classification results of static images using a Caltech-256 object category dataset demonstrate that our method can visualize spatial relations of categories while maintaining time-series characteristics. Moreover, we emphasize the effectiveness of our method for category formation of appearance changes of objects.

1 Introduction

Because of the advanced progress of computer technologies and machine learning algorithms, generic object recognition has been studied actively in the field of computer vision [11]. Generic object recognition is defined as a capability by which a computer can recognize objects or scenes to their general names in real images with no restrictions, i.e., recognition of category names from objects or scenes in images. In actual environments, the number of categories is mostly unknown. Moreover, the categories are not known uniformly.

Learning-based object classification methods are roughly divisible into supervised object classification methods and unsupervised object classification methods. Supervised object classification methods require training datasets including teaching signals extracted from ground-truth labels. In contrast, unsupervised object classification methods require no teaching signals. Studies of unsupervised object classification methods have been active gradually. The subject has attracted attention because it might provide technologies to classify visual information flexibly in various environments.

As unsupervised object classification methods, Sivic et al. proposed a method using pLSA and LDA, which are generative models from the statistical text literature [8]. They modeled an image including instances of several categories

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 427–434, 2011.

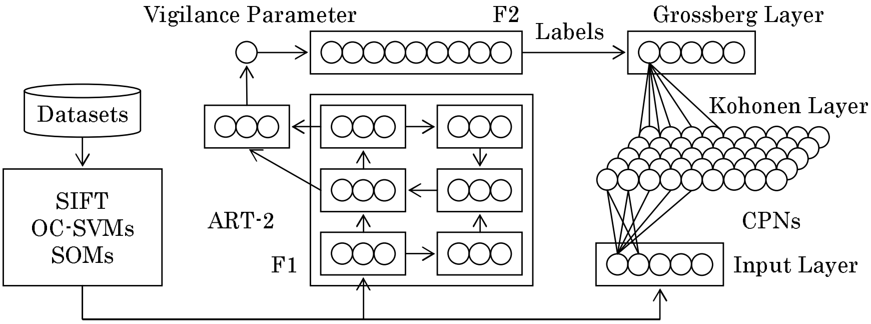


Fig. 1. Whole architecture of our method

as a mixture of topics and attempted to discover topics as object categories from numerous images. Zhu et al. proposed Probabilistic Object Models (POMs) that improved their method and enabled classification, segmentation, and recognition of objects [2]. Todorovic et al. proposed an unsupervised identification method using optical, geometric, and topological characteristics of multiscale regions consisting of two-dimensional objects [10]. They represented each image as a tree structure by division of multi-scale images. However, these methods include the restriction of prior settings of the number of classification categories. Therefore, these methods are applied only slightly to classification problems in an actual environment for which the number of categories is unknown.

This paper presents unsupervised feature selection and category formation without previous setting of the number of categories. Our method has the following four features. First, our method can localize target feature points using One Class-Support Vector Machines (OC-SVMs) [7] without previous setting of boundary information. Second, our method can generate labels as a candidate of categories for input images while maintaining stability and plasticity together. Third, automatic labeling of category maps can be realized using labels created using Adaptive Resonance Theory-2 (ART-2) as teaching signals for Counter Propagation Networks (CPNs). Fourth, our method can present the diversity of appearance changes for visualizing spatial relations of each category on a two-dimensional map of CPNs. Through object classification experiments, we evaluate our method using the Caltech-256 object category dataset [4].

2 Proposed Method

In generic object recognition, it is a challenging task to develop a unified model to address all steps from feature representation to creation of classifiers. The aim of our study is the realization of category formation for generic object recognition to apply theories with different characteristics for each step. Fig. 1 depicts the network architecture of our method. The procedures are the following.

1. Selecting feature points using OC-SVMs
2. Creating visual words using Self-Organizing Maps (SOMs)
3. Generating labels using ART-2
4. Creating a category map using CPNs

Procedures 1. through 3., which correspond to preprocessing, are based on the representation of Bag-of-Features (BoF) [3]. We apply OC-SVMs to select feature points for localizing target regions in an image. For producing visual words, we use SOMs, which can learn neighborhood regions while updating the cluster structure, although k-means must determined data of the center of a cluster. Actually, SOMs can represent visual words that minimize misclassification [9]. Furthermore, the combination of ART-2 and CPNs enables unsupervised category formation that labels a large quantity of images in each category automatically.

2.1 Selected Feature Points with OC-SVMs

As described earlier, the OC-SVMs are unsupervised learning classifiers that estimate the dense region without using density functions [7]. Our target is Scale-Invariant Feature Transform (SIFT) feature points on an object for recognition. Therefore, target regions and target feature points respectively mean object regions and feature points on an object. The OC-SVMs are unsupervised-learning-based binary classifiers that enable density estimation without estimating a density function. Therefore, OC-SVMs can apply to real-world images without boundary information.

The discriminant function $f(\cdot)$ is calculated to divide input feature vectors x_{i} into two parts. The position of the hyperplane is changed according to parameter ν, which controls outliers of input data with change, and which has range of 0–1.

$$f(x) = sgn(\omega^\top \varPhi(x) - \rho). \tag{1}$$

Here, ω and ρ ($\rho \in R$) represent a coefficient and a margin. Therein, z_{i} represents results of x_{i} to the high-dimension feature space.

$$\varPhi : x_i \mapsto z_i \tag{2}$$

The restriction is set to the following.

$$\omega^\top z_i \geq \rho - \zeta_i, \zeta_i \geq 0, 0 < \nu \leq 1 \tag{3}$$

Here, ζ represents relaxation variable vectors. The optimization problem is solved with the following restriction

$$\frac{1}{2}\|\omega\|^2 + \frac{1}{\nu l}\sum_{i=1}^{l} \zeta_i - \rho \rightarrow \min \omega, \zeta, \text{and } \rho \tag{4}$$

Parameter ν of OC-SVMs is a high limit of unselected data and lower limit of support vectors if the solution of the optimization problem (4) fulfills $\rho \neq 0$.

2.2 Creating Visual Words with SOMs

For our method, we apply SOMs, not k-means, which is generally used in BoF, for creating visual words. In the learning step, SOMs update weights while maintaining topological structures of input data. Actually, SOMs create neighborhood regions around the burst unit, which demands a response of the input data. Therefore, SOMs can classify various data whose distribution resembles the training data. In addition, Terashima et al. reported that SOMs are superior to k-means as an unsupervised classification method that is useful to minimize misrecognition [9]. The learning algorithm of SOMs [6] is the same as the algorithm used between the Input-Kohonen layers of CPNs. In this method, we used all SIFT features for creating visual words at the learning step of SOMs. We used SIFT features selected by OC-SVMs for generating histograms based on visual words. Based on our preliminary experiment, we set the learning iteration to 100,000 times. Additionally, we set the number of units of the Kohonen layer to 100 units. We created visual words to extract weights between Kohonen layer units and input layer units.

2.3 Generating of Labels with ART-2

Actually, ART-2 [1] is a theoretical model of unsupervised neural networks of incremental learning that forms categories adaptively while maintaining stability and plasticity together. Features of time-series images from the mobile robot change with time. Using ART-2, our method enables an unsupervised category formation that requires no setting of the number of categories.

The learning algorithm of ART-2 is the following.

1) Input data x_i are presented to the F1.
2) Search for the maximum active unit T_j as

$$T_J(t) = max(\sum_j p_i(t) Z_{ij}(t)). \tag{5}$$

3) Top-down weights Z_{ji} and bottom-up weights Z_{ij} are updated as

$$\frac{d}{dt} Z_{Ji}(t) = d[p_i(t) - Z_{Ji}(t)], \tag{6}$$

$$\frac{d}{dt} Z_{iJ}(t) = d[p_i(t) - Z_{iJ}(t)]. \tag{7}$$

4) The vigilance threshold ρ is used to judge whether input data correctly belong to a category.

$$\frac{\rho}{e + \|r\|} > 1, \quad r_i(t) = \frac{u_i(t) + cp_i(t)}{e + \|u\| + \|cp\|}. \tag{8}$$

Here, p_i and u_i are sublayers on F1. When (8) is true, the active units reset and return 2) to search again. Repeat 1) and 4) until the rate of change of F1 is sufficiently small if (8) is not true.

2.4 Creating Category Maps with CPNs

The CPN [5] actualizes mapping and labeling together. Such networks comprise three layers: an input layer, a Kohonen layer, and a Grossberg layer. In addition, CPNs learn topological relations of input data for mapping weights between units of the input-Kohonen layers. The resultant category formations are represented as a category map on the Kohonen layer. Our method can reduce these labels using the Winner-Takes-All competition of CPNs. In addition, our method can visualize relations between categories on the category map of CPNs. Detailed algorithms of ART-2 and CPNs are the following.

The CPN learning algorithm is the following. $u^i_{n,m}(t)$ are weights from an input layer unit $i (i = 1, ..., I)$ to a Kohonen layer unit $(n, m)(n = 1, ..., N, m = 1, ..., M)$ at time t. Therein, $v^j_{n,m}(t)$ are weights from a Grossberg layer unit j to a Kohonen layer unit (n, m) at time t. These weights are initialized randomly. The training data $x_i(t)$ show input layer units i at time t. The Euclidean distance $d_{n,m}$ separating $x_i(t)$ and $u^i_{n,m}(t)$ is calculated as

$$d_{n,m} = \sqrt{\sum_{i=1}^{I}(x_i(t) - u^i_{n,m}(t))^2}. \tag{9}$$

The unit for which $d_{n,m}$ is smallest is defined as the winner unit c as

$$c = argmin(d_{n,m}). \tag{10}$$

Here, $N_c(t)$ is a neighborhood region around the winner unit c. $u^i_{n,m}(t)$ and $v^j_{n,m}(t)$ of $N_c(t)$ are updated as

$$u^i_{n,m}(t + 1) = u^i_{n,m}(t) + \alpha(t)(x_i(t) - u^i_{n,m}(t)). \tag{11}$$

$$v^j_{n,m}(t + 1) = v^j_{n,m}(t) + \beta(t)(t_j(t) - v^j_{n,m}(t)). \tag{12}$$

In that equation, $t_j(t)$ is the teaching signal to be supplied to the Grossberg layer. Furthermore, $\alpha(t)$ and $\beta(t)$ are the learning rate coefficients that decrease with the progress of learning. The learning of CPNs repeats up to the learning iteration that was set previously.

3 Experimental Results Obtained Using the Caltech-256 Dataset

The target of this experiment is object classification of static images because Caltech-256 [4] has no temporal factors in each category. We use the highest 20 categories with the number of images in 256 categories. The results of selection of SIFT features and recognition accuracy for classification of 5, 10, and 20 categories are the following.

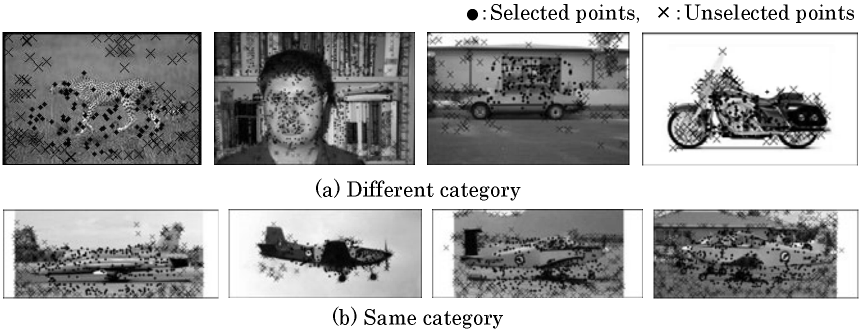


Fig. 2. Results of selected SIFT feature points in the same and different categories on Caltech-256

3.1 Category Formation Results

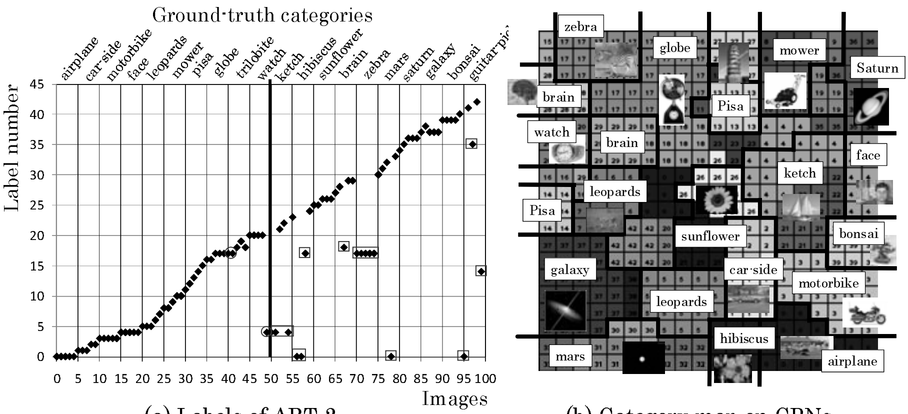
Figure 2 depicts results of selected feature points using OC-SVMs on eight sample images of Caltech-256. Fig. 2 (a) shows that our method can select feature points of target objects in images of different categories. In addition, Fig. 2 (b) shows that our method can select feature points around the wings that characterize airplanes for various images of the Airplane category.

Figure 3 (a) depicts labels by ART-2 on 20-object classification. The vertical and horizontal axes respectively represent labels and images. The bold line shows the number of images in 10 categories. The circles and squares portray images for which ART-2 confused labels on 10 and 20 categories, respectively. In the 10-object classification, ART-2 generated independent labels in all categories, although three images of two labels are confused. In the 20-object classification, independent labels of 19 categories are generated, except for the Zebra category that is confused of all images, although 16 images of five labels are confused. Confusion of labels occurs often in images of Ketch, Hibiscus, and Guitar-pick categories. Although confused labels are restrained until 10-object classification, numerous confused labels are apparent in the 20-object classification.

Figure 3 (b) depicts a category map generated with CPNs on 20-object classification. The names of categories and the number of images are shown on the category map. For all images in each category, 11 categories are mapped to neighborhood units. The CPNs created categories for mapping neighborhood units on the category map in images of each category by which ART-2 generated several labels. In addition, categories without their names are mapped images of different categories.

3.2 Recognition Accuracy

Table 1 portrays results of recognition accuracy obtained using our method and POMs, the existing state-of-the-art unsupervised classification method, as re-



(a) Labels of ART-2 (b) Category map on CPNs

Fig. 3. Results of formed labels using ART-2 and category map of 20 categories

Table 1. Recognition accuracy for learning and testing datasets used in Caltech-256

Number of categories	Our method		POMs by Chen et al. [2]	
	Learning	Testing	Learning	Testing
5	96%	76%	68%	73%
10	94%	42%	75%	72%
20	81%	45%	–	–
26	–	–	77%	67%

ported by Chen et al. [2]. The recognition accuracy values obtained using our method were, respectively, 96%, 94%, and 81% for training datasets and 76%, 42%, and 45% for testing datasets in 5, 10, and 20 categories. The recognition accuracy values obtained using POMs were, respectively, 68%, 75%, and 77% for training datasets and 73%, 72%, and 67% for testing datasets in 5, 10, and 26 categories. In five-category classification, the recognition accuracy of our method is higher than that of POMs for both training and testing datasets. In results for more than 10 categories, the recognition accuracy of our method was lower than that of POMs for the testing dataset, but the recognition accuracy of our method was higher than that of POMs for the training dataset. We consider that the result of our method is inferior to POM results because of over-fitting. We used the category map of fixed size, 20 × 20 unit, for all recognition targets. For improving expression and mapping capabilities of CPNS, we will consider introduction of a mechanism to change a suitable size of the category map according to the number of categories to be classified.

Actually, objects of various types exist in an actual environment. In our daily life, it is almost unknown how many objects exist in a room. Therefore, it is unrealistic to present the number of categories in advance. POMs require setting of a number of categories in advance. Our method can classify objects without prior setting of the number of categories. Therefore, our method is effective for application to problems that are known as challenging tasks of classification of categories whose ranges and types are unclear.

4 Conclusion

This paper presented an unsupervised method of SIFT feature points selection using OC-SVMs and category formation combined with incremental learning of ART-2 and self-mapping characteristic of CPNs. Our method enables feature representation that contributes to improved accuracy of classification for selecting feature points to concentrate characterized information of an image. Moreover, our method can visualize spatial relations of labels and integrate redundant and similar labels generated with ART-2 as a category map using self-mapping characteristics and neighborhood learning of CPNs. Therefore, our method can represent diverse categories.

Future studies must be conducted to develop methods to extract boundaries among clusters automatically and to determine a suitable number of categories from category maps of CPNs.

References

1. Carpenter, G.A., Grossberg, S.: Art 2: Stable self-organization of pattern recognition codes for analog input patterns. Applied Optics 26
2. Chen, Y., Zhu, L., Yuille, A., Zhang, H.
3. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Proc. European Conf. Computer Vision
4. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset, California Institute of Technology Technical Report
5. Hecht-Nielsen, R.: Counterpropagation networks. Applied Optics 26(23), 4979–4984 (1987)
6. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics 43(1), 59–69 (1982)
7. Scholkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high dimensional distribution. Neural Computation 13, 1443–1471 (2001)
8. Sivic, J., Russell, B.C., Zisserman, A., Efros, A.A., Freeman, W.T.: Discovering objects and their localization in images. In: Proc. Conf. Computer Vision
9. Terashima, M., Shiratani, F., Yamamoto, K.: Unsupervised cluster segmentation method using data density histogram on self-organizing feature map. Journal of the Institute of Electronics, Information, and Communication Engineers (D–II), 1280–1290 (1996)
10. Todorovic, S., Ahuja, N.: Unsupervised category, modeling, recognition, and segmentation in images. IEEE Trans. Pattern Analysis and Machine Intelligence 30(12), 2158–2174 (2008)
11. Yanai, K.: The current state and future directions on generic object recognition. Journal of Information Processing: The Computer Vision and Image Media 48

Object Recognition with the HOSVD of the Multi-model Space-Variant Pattern Tensors

Bogusław Cyganek

AGH University of Science and Technology,
Al. Mickiewicza 30, 30-059 Kraków, Poland
cyganek@agh.edu.pl

Abstract. The paper presents a framework for object recognition with the multi-model space-variant approach in the log-polar domain built into the multilinear tensor classifier. Thanks to this the method allows recognition of rotated and/or scaled objects taking advantage of the foveal and peripheral information. Recognition is done in the multilinear subspaces obtained after the higher-order singular value decomposition of the pattern tensor. The experiments show high accuracy and robustness of the proposed method.

1 Introduction

Object recognition is one of the fundamental tasks of Computer Vision and also one of the most demanding ones due to discrepancy between a variety of real objects and their images obtained under different conditions such as viewpoints, illuminations, image resolutions, noise, etc. In this context the methods employing tensor decompositions show very high accuracy and robustness [11][6]. A method based on higher-order SVD (HOSVD) for digit classification was proposed by Savas *et al.* [9]. One of its main properties is robustness to spatial variations of patterns. However, to cope with an inherent rotation and scale the special means have to be undertaken. For instance, in a system for road signs recognition, to account for object rotations the rotated versions of the test patterns are generated which then compose the deformed prototypes tensor (DPT) [4]. To cope with scale variations, also different versions of scaled object can be generated or a test object before the classification needs to be registered to the known reference dimensions [2]. All these increase tensor size and for some distorted test objects do not guarantee sufficient accuracy. On the other hand, it is well known that the transformation from the Euclidean to the log-polar (LP) representation maps rotation and scale into linear shifts. However, linear shifts can be easily accommodated by the HOSVD of the DPT [4]. Nevertheless, LP belongs to the group of space variant transformations, i.e. it is not translation invariant. In other words, features describing a target depend on a viewpoint. If many different viewing positions are used to prepare the prototype patterns then the multiple prototype models is obtained. This was proposed by Traver *et al.* [10] as a multiple-model approach (MMA). MMA has many beneficial properties especially in the context of active vision systems. During our experiments with the HOSVD it was observed that the MMA fits well into the HOSVD framework since each of its subspaces is built

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 435–442, 2011.

separately around each set of prototypes of a single training class. Thus, the main contribution of this paper is incorporation of the MMA into the recognition framework with the HOSVD and DPT. The method takes advantage of a decreasing resolution in the representation going from foveal center toward the view periphery. The method appears to be very robust in terms of accuracy and operation time. It was checked in the prototypical automotive application of a driver assisting system which contains the road signs recognition, as well as driver inattention monitoring system based on a driver's eye recognition. Rest of the paper is organized as follows: Section 2 presents the concept of the multiple spatial model in the log-polar space. Section 3 deals with MMA built into the HOSVD framework. Experimental results are presented in 4. The paper ends with conclusions in Section 5.

2 Multiple Spatial Log-Polar Models

Following the works by Jurie [5] and Traver *et al.* [10], the multiple spatial log-polar model is defined as a set of different image positions $\{P_i\}$, for which a separate set of features $\{F_i\}$ is computed, describing a target as viewed at $\{P_i\}$. In our case $\{P_i\}$ is proposed to cover the foveal and peripheral areas of the image space, while $\{F_i\}$ is the set of log-polar representation of the target T_i.

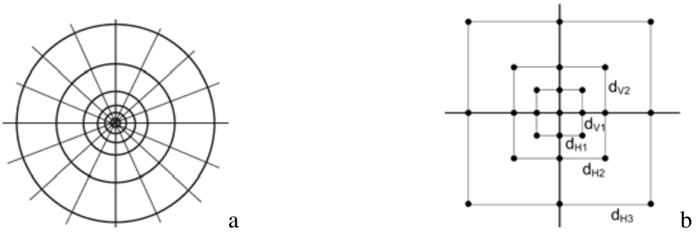


Fig. 1. The log-polar space (a). Points on a rectangular grid defining multiple spatial models (b).

Fig. 1a depicts the log-polar mapping while Fig. 1b depicts the set of $\{P_i\}$ points at which the log-polar representations $\{F_i\}$ are computed. The distances d_{Hi} and d_{Vi} were chosen as 1, 2, and 4, respectively. The nonlinear LP transformation with a center $\mathbf{c}=(c_1,c_2)\in\{P_i\}$ transforms a point $\mathbf{x}=(x_1,x_2)$ into $\mathbf{y}=(r,\varphi)$, as follows

$$r=\log_B\left(\sqrt{\left(x_1-c_1\right)^2+\left(x_2-c_2\right)^2}\right),\ \varphi=\arctan\left(\left(x_2-c_2\right)/\left(x_1-c_1\right)\right), \tag{1}$$

where B is a base of a logarithm which should be greater than 1.0. Usually B is chosen to fit the value of r_{max} which is the maximal distance from the center $\mathbf{c}$ to the points of the original image. In this case it is given as follows

$$B=\exp\left[\ln\left(d_{\max}\right)/r_{\max}\right], \tag{2}$$

where $d_{max}>1$, $r_{max}>1$, and $d_{max}=min\{d_{1max}, d_{2max}\}$ is a minimal distance from a chosen center C and surrounding window, r_{max} is the maximal number of columns in the LP image.

3 Object Recognition with the Higher-Order Singular Value Decomposition and Multiple Log-Polar Models

Tensors allow description of physical laws which transform appropriately with a change of the coordinate system. These can be also seen as the multidimensional arrays of data. Thus, scalars, vectors, and matrices all are tensors. Also images in different formats can be seen as tensors. This allows an explicit control of intrinsic dimensions, but also with help of tensor decompositions - an insight into intrinsic information hidden in massive amount of pixels.

In this work a tensor is composed of space-variant log-polar representations of the prototype models. This way a set of DPTs is obtained, one for each pattern with its LP representations. A number of these representations can be different for each prototype. Then each DPT is decomposed with the HOSVD, which allows construction of a space spanned by the dominating base tensors. Recognition is done checking distances of a test object projected into these spaces. In this section we provide a brief overview of tensors and their HOSVD decomposition in the context of pattern recognition in computer vision. More on this can be found in literature [6][7].

3.1 Basic Concepts of Tensor Algebra

When processing tensors, the first important concepts is tensor flattening. For an *P-th* order tensor $\mathcal{T} \in \Re^{N_1 \times N_2 \times \ldots N_P}$ the *k-mode vector* (or *a fiber*) of $\mathcal{T}$ is defined as a vector obtained from the elements of $\mathcal{T}$ by varying only one index n_k when keeping all other fixed. If from $\mathcal{T}$ a following matrix

$$\mathbf{T}_{(k)} \in \Re^{N_k \times \left(N_1 N_2 \ldots N_{k-1} N_{k+1} \ldots N_P\right)} \tag{3}$$

is formed, then columns of $\mathbf{T}_{(k)}$ are *k-mode* vectors of $\mathcal{T}$. The k-mode representation of a tensor is obtained by selecting the k-th index to become a row index of its flatten representation. On the other hand its column index is a product of all other P-1 indices. Nevertheless, where an element of the tensor is stored in memory depends on an assumed permutation of these P-1 indices, which gives (P-1)! possibilities. From these only two, i.e. forward and backward cycle modes, are used [6]..

The second important concept is a k-mode multiplication of a tensor $\mathcal{T} \in \Re^{N_1 \times N_2 \times \ldots N_P}$ and a matrix $\mathbf{M} \in \Re^{Q \times N_k}$. In result a following tensor $\mathcal{S} \in \Re^{N_1 \times N_2 \times \ldots N_{k-1} \times Q \times N_{k+1} \times \ldots N_P}$ is obtained whose elements can be expressed as

$$\mathcal{S}_{n_1 n_2 \ldots n_{k-1} q n_{k+1} \ldots n_P} = \left(\mathcal{T} \times_k \mathbf{M}\right)_{n_1 n_2 \ldots n_{k-1} q n_{k+1} \ldots n_P} = \sum_{n_k=1}^{N_k} t_{n_1 n_2 \ldots n_{k-1} n_k n_{k+1} \ldots n_P} m_{q n_k} . \tag{4}$$

The third important concept is the HOSVD which is an analog to the SVD for matrices [6][7]. HOSVD allows any P-dimensional tensor $\mathcal{T} \in \Re^{N_1 \times N_2 \times \ldots N_m \times \ldots N_n \times \ldots N_P}$ to be equivalently represented as follows

$$\mathcal{T} = \mathcal{Z} \times_1 \mathbf{S}_1 \times_2 \mathbf{S}_2 \ldots \times_P \mathbf{S}_P . \tag{5}$$

$\mathbf{S}_k$ are unitary matrices of dimensions $N_k \times N_k$, called *mode matrices.* $\mathcal{Z} \in \Re^{N_1 \times N_2 \times \ldots N_m \times \ldots N_n \times \ldots N_P}$ is *a core tensor* which fulfills the two properties [6][7]:

1. Two subtensors $\mathcal{Z}_{n_k=a}$ and $\mathcal{Z}_{n_k=b}$, are orthogonal for all possible values of k for which $a \neq b$, i.e.

$$\mathcal{Z}_{n_k=a} \cdot \mathcal{Z}_{n_k=b} = 0 , \tag{6}$$

2. All subtensors of $\mathcal{Z}$ for all k can be ordered according to their Frobenius norms

$$\left\| \mathcal{Z}_{n_k=1} \right\| \geq \left\| \mathcal{Z}_{n_k=2} \right\| \geq \ldots \geq \left\| \mathcal{Z}_{n_k=N_P} \right\| \geq 0 , \tag{7}$$

Finally, the *a-mode* singular value of $\mathcal{T}$ is defined as follows

$$\left\| \mathcal{Z}_{n_k=a} \right\| = \sigma_a^k . \tag{8}$$

3.2 Pattern Recognition with the Deformable Prototype Tensor

For each mode matrix $\mathbf{S}_i$ in (5) the following sum can be constructed

$$\mathcal{T} = \sum_{h=1}^{N_P} \mathcal{T}_h \times_P \mathbf{s}_P^h , \tag{9}$$

thanks to the commutative properties of the k-mode multiplication. In the above

$$\mathcal{T}_h = \mathcal{Z} \times_1 \mathbf{S}_1 \times_2 \mathbf{S}_2 \ldots \times_{P-1} \mathbf{S}_{P-1} \tag{10}$$

denote the basis tensors and $\mathbf{s}^{\mathrm{h}}_{\mathrm{P}}$ are columns of the unitary matrix $\mathbf{S}_P$. Since $\mathcal{T}_{\mathrm{h}}$ is of dimension P-1 then $\times_P$ in (9) is an outer product, i.e. a product of two tensors of dimensions P-1 and 1. Let us now observe that due to the orthogonality properties of the core tensor $\mathcal{Z}$ in (10), $\mathcal{T}_{\mathrm{h}}$ are also orthogonal. Thus, they can constitute a basis.

In this space pattern recognition with HOSVD can be stated as testing a distance of a given test pattern $\mathbf{P}_x$ to its projections in each of the spaces spanned by the set of the bases $\mathcal{T}_{\mathrm{h}}$ in (10). This can be expressed as the following minimization problem

$$\min_{i, c_h^i} \underbrace{\left\| \mathbf{P}_x - \sum_{h=1}^{N} c_h^i \mathcal{T}_h^i \right\|^2}_{Q_i} , \tag{11}$$

where the scalars $c^i{}_h$ denote unknown coordinates of $\mathbf{P}_x$ in the space spanned by $\mathcal{T}_h^i$, and $N \leq N_P$ denotes a number of chosen dominating components.

To solve (11) the squared norm Q of (11) is created for a selected i, as follows

$$Q = \left\| \mathbf{P}_x - \sum_{h=1}^{N} c_h \mathcal{T}_h \right\|^2 = \left\langle \mathbf{P}_x - \sum_{h=1}^{N} c_h \mathcal{T}_h , \mathbf{P}_x - \sum_{h=1}^{N} c_h \mathcal{T}_h \right\rangle \tag{12}$$

Now, to find a minimum of (11) for each c$_h$, the set of derivatives with respect to each c$_h$ is computed and then equated to 0. This leads to the following expression

$$c_h = \left\langle \mathcal{T}_h , \mathbf{P}_x \right\rangle \Big/ \left\langle \mathcal{T}_h , \mathcal{T}_h \right\rangle, \tag{13}$$

Now, for each i, (13) can be back substituted into (11), which yields the following residual

$$\rho_i = \left\| P_x - \sum_{h=1}^{N} \frac{\left\langle \mathcal{T}_h^i , \mathbf{P}_x \right\rangle}{\left\langle \mathcal{T}_h^i , \mathcal{T}_h^i \right\rangle} \mathcal{T}_h^i \right\|^2 . \tag{14}$$

Assuming further that $\mathcal{T}_h^i$ and $\mathbf{P}_x$ are normalized the following is obtained (the *hat* notation relates to the normalized tensors)

$$\rho_i = 1 - \sum_{h=1}^{N} \left\langle \hat{\mathcal{T}}_h^i , \hat{\mathbf{P}}_x \right\rangle^2 . \tag{15}$$

Thus, to minimize (11) we need to maximize the following value

$$\hat{\rho}_i = \sum_{h=1}^{H} \left\langle \hat{\mathcal{T}}_h^i , \hat{P}_x \right\rangle^2 , \tag{16}$$

In other words, our system returns a class i for which the corresponding ρ_i from (16) is the largest.

4 Experimental Results

All layers of the system were implemented in C++ using the HIL library [3]. The experimental setup consists of the computer with 8 GB RAM and Pentium® Core Q 820 (clock 1.73 GHz). Two types of objects were tested for recognition, facial regions and the road signs, which training databases are depicted in Fig. 2a and Fig. 2b, respectively. The prototype patterns are transformed into the LP-MMA from which the pattern tensor $\mathcal{T}$is constructed. Then, HOSVD is computed from $\mathcal{T}$, from which the sets of base tensors $\mathcal{T}_h^i$ are obtained for each pattern separately in accordance with (10), as described in the previous sections.

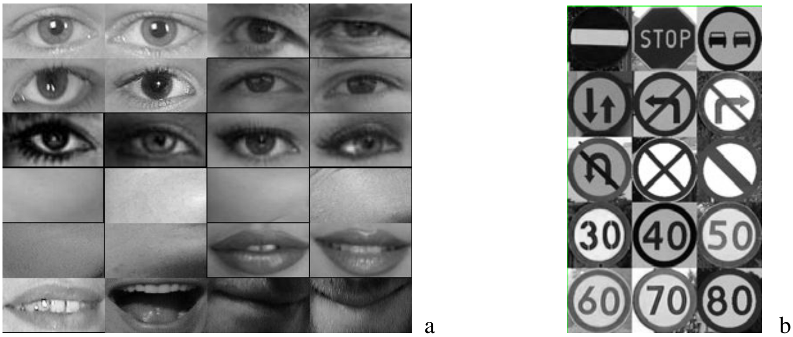


Fig. 2. The two databases used in the experiments. Face regions database (a), road signs (b).

Fig. 3 depicts first five tensors $\mathcal{T}_h$ for the speed limit sign "*70 km/h*" from the database in Fig. 2b. The corresponding subtensors $\mathcal{Z}_n$, which correspond to the "energy" factor (8), spread out from the left top corner to the bottom right one [6].

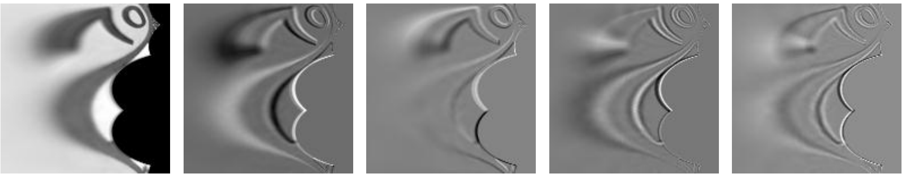


Fig. 3. First five tensors $\mathcal{T}_h$ of the "70 km/h" speed limit sign

During operation the HOSVD classifier selects a class of corresponding to the best subspace. That is, the subtensors $\mathcal{T}_h^i$ are used to compute distances (16) of the test pattern $\mathbf{P}_{\mathrm{x}}$ to all subspaces spanned by $\mathcal{T}_h^i$.

Fig. 4 depicts stages of processing of real traffic scenes from which the sign areas are cropped and their LP representations computed. These are then fed to the HOSVD-MMA classifier. The measured accuracy is 94% on average, thus it is better than in our previous system [2] and comparable with the system operating exclusively in the spatial domain and the HOSVD, presented in [4]. However, the proposed method performs better than [4] in the case of imprecisely cropped objects, thanks to the employed multiple model approach.

Fig. 5 depicts results of operation of our method on three selected frames showing a person, shown in the first column. The second column of Fig. 5 contains the skin segmented areas (in white). Third column shows compact skin regions detected with the adaptively growing window method, described in [2].

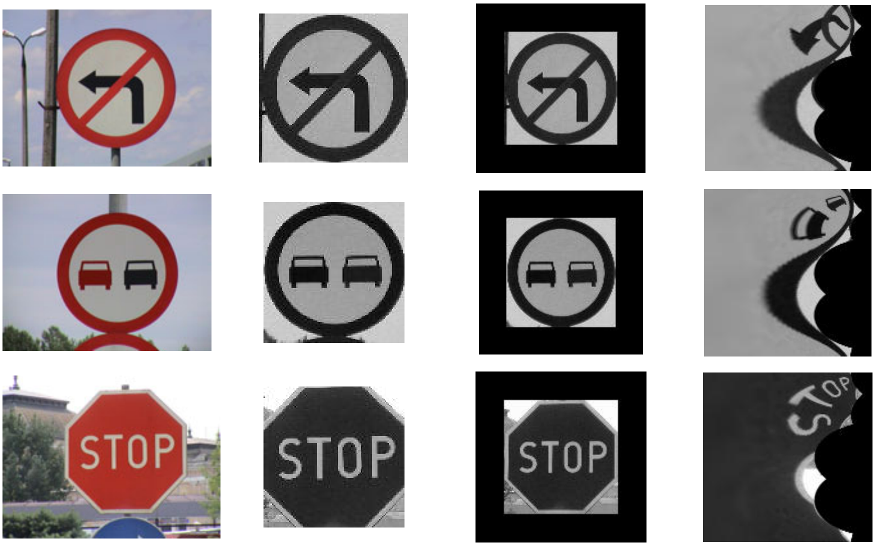


Fig. 4. Stages of processing of real images to their LP representations. Original image (first column), cropped area of interest based on color information (second column), the framed patterns (third column), the log-polar version of the test pattern (fourth column).

Finally, fourth column of Fig. 5 depicts correctly identified eye regions. Experiments were conducted on a database of selected test images containing persons with well visible faces in good lighting conditions (i.e. daily or artificial light).

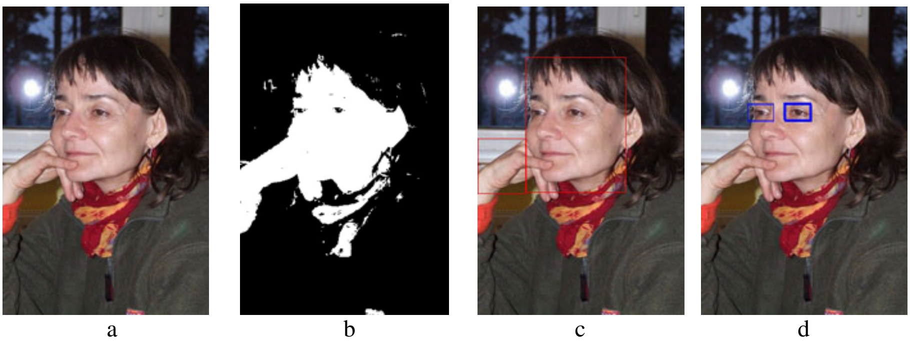


Fig. 5. Results of eye recognition for three images. Original image (a). Skin binary map (b). Compact skin regions detected with the adaptive window growing algorithm (c). Detected eyes with the HOSVD classifier trained with database in Fig. 2a (d).

To test accuracy of the method it was compared with answers of a human operator. The resulting average accuracy is 97%. This compares favorably with the reported results [12][1]. Examination of the misclassified cases reveals that problems are usually due to wrong initial skin segmentation. More precise segmentation offer other methods [8][12]. However, they require more computational effort than the used one.

The average execution time for an 640x480 RGB image is in order of 150-200 ms, which allows real time operation.

5 Conclusions

In this paper we propose a novel method for object recognition which connects the multilinear HOSVD based classification and the multiple model approach operating in the space variant log-polar space. Such connection allows recognition of rotated or scaled objects and shows better performance than a single HOSVD. The experiments were performed on two types of objects for different applications. The first is recognition of the road signs. The second is recognition of human eyes. The obtained results show high accuracy and fast operation which allows real time processing even in software implementation.

Acknowledgement. This work was supported from the Polish funds for scientific research in 2011.

References

1. Chiang, C.-C., Tai, W.-K., Yang, M.-T., Huang, Y.-T., Huang, C.-J.: A novel method for detecting lips, eyes and faces in real time. Real-Time Imaging 9, 277–287 (2003)
2. Cyganek, B.: Circular Road Signs Recognition with Soft Classifiers. Integrated Computer-Aided Engineering 14(4), 323–343 (2007)
3. Cyganek, B., Siebert, J.P.: An Introduction to 3D Computer Vision Techniques and Algorithms. Wiley, Chichester (2009)
4. Cyganek, B.: An Analysis of the Road Signs Classification Based on the Higher-Order Singular Value Decomposition of the Deformable Pattern Tensors. In: Blanc-Talon, J., Bone, D., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2010, Part II. LNCS, vol. 6475, pp. 191–202. Springer, Heidelberg (2010)
5. Jurie, F.: A new log-polar mapping for space variant imaging. Application to face detection and tracking. Pattern Recognition 32, 865–875 (1999)
6. de Lathauwer, L.: Signal Processing Based on Multilinear Algebra. PhD dissertation, Katholieke Universiteit Leuven (1997)
7. de Lathauwer, L., Moor de, B., Vandewalle, J.: A Multilinear Singular Value Decomposition. SIAM Journal of Matrix Analysis and Applications 21(4), 1253–1278 (2000)
8. D'Orazio, T., Leo, M., Guaragnella, C., Distante, A.: A visual approach for driver inattention detection. Pattern Recognition 40, 2341–2355 (2007)
9. Savas, B., Eldén, L.: Handwritten digit classification using higher order singular value decomposition. Pattern Recognition 40, 993–1003 (2007)
10. Traver, V.J., Bernardino, A., Moreno, P., Santos-Victor, J.: Appearance-based object detection in space-variant images: A multi-model approach. In: Campilho, A., Kamel, M. (eds.) ICIAR 2004. LNCS, vol. 3211, pp. 538–546. Springer, Heidelberg (2004)
11. Vasilescu, M.A.O., Terzopoulos, D.: Multilinear analysis of image ensembles: TensorFaces. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 447–460. Springer, Heidelberg (2002)
12. Zhu, Z., Jib, Q.: Robust real-time eye detection and tracking under variable lighting conditions and various face orientations. Computer Vision and Image Understanding 98, 124–154 (2005)

Precise Eye Detection Using Discriminating HOG Features

Shuo Chen and Chengjun Liu

Department of Computer Science, New Jersey Institute of Technology,
323 M.L. King Boulevard, University Height, Newark, NJ 07032, USA
{sc77,chengjun.liu}@njit.edu

Abstract. We present in this paper a precise eye detection method using Discriminating Histograms of Oriented Gradients (DHOG) features. The DHOG feature extraction starts with a Principal Component Analysis (PCA) followed by a whitening transformation on the standard HOG feature space. A discriminant analysis is then performed on the reduced feature space. A set of basis vectors, based on the novel definition of the within-class and between-class scatter vectors and a new criterion vector, is defined through this analysis. The DHOG features are derived in the subspace spanned by these basis vectors. Experiments on Face Recognition Grand Challenge (FRGC) show that (i) DHOG features enhance the discriminating power of HOG features and (ii) our eye detection method outperforms existing methods.

Keywords: Histograms of Oriented Gradients (HOG), Discriminant Analysis, Eye Detection, Face Recognition Grand Challenge (FRGC).

1 Introduction

Eye detection has a significant impact on the performance of face recognition due to the *Curse of Alignment* [13]. Even a slight detection error (e.g., 5 pixels) will dramatically reduce the face recognition performance [13], [12]. Detecting eyes in images is a challenging task due to the wide change of face pose and expressions (e.g., closed eyes) and the various obstructions (e.g., glasses and hats). Many eye detection methods have been proposed over the last decade [7] [10] [15] [13] [14] [4]. However, many problems, especially in detection accuracy, still exist.

Recently, Dalal & Triggs [3] presented the Histograms of Oriented Gradients (HOG) for human detection and got excellent detection performance. The basis idea of HOG features is that local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions. In this paper, we present a novel discriminating HOG (DHOG) features to reduce the dimensionality of the standard HOG features while enhance its discriminating power. The most widely used dimensionality reduction technique is probably the Principal Component Analysis (PCA) [5]. Although PCA can derive the optimal representing features, it can not derive the optimal discriminating features. An alternative is the Fisher Linear Discriminant (FLD)

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 443–450, 2011.

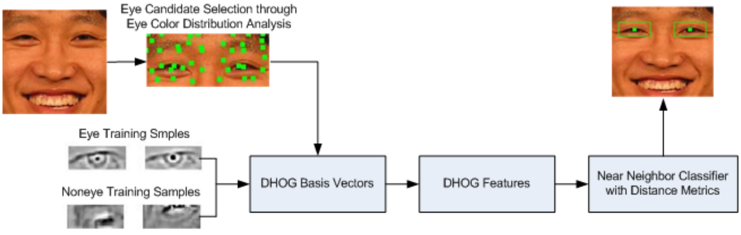


Fig. 1. Work flow chart of our eye detection method

[5]. For any L-class pattern classification problem, FLD derives a compact and well-separated features with the dimensionality of $L-1$. However, when applied to the two-class detection problem, FLD only derives one feature, which will lead to the significant loss of data information and a very poor classification performance.

Our DHOG feature extraction starts with a Principal Component Analysis (PCA) followed by a whitening transformation on the HOG feature space. A discriminant analysis is then performed on the reduced feature space. A set of basis vectors, based on the novel definition of the within-class and between-class scatter vectors and a new criterion vector, is defined through this analysis. The DHOG features are then derived in the subspace spanned by these basis vectors.

Our eye detection method is shown is Fig. 1. First, a face is detected using the BDF method proposed in [8] and normalized to the size of 128×128. Then geometric constraints are applied to localize the eyes, which means eyes are only searched in the top half of the detected face. Then the eye detection is achieved by two steps: the feature based eye candidate selection and appearance based validation. The selection stage chooses eye candidates through an eye color distribution analysis in the YCbCr color space based on the observation that the pixels in the eye region, compared with other skin area, have higher chrominance blue (Cb) value, lower chrominance red (Cr) value, and lower luminance (Y) value [2]. The validation stage first extracts the DHOG features of each candidate and then a near neighbor classifier with the distance metric is applied for classification to detect the center of the eye among these candidates. Usually, there are multiple eyes detected around the pupil center. The final eye location is the average of these multiple detections.

We perform the experiments on Face Recognition Grand Challenge (FRGC) database to evaluate the performance of DHOG features and our eye detection method. Experiment results show that (i) DHOG features enhance the discriminating power of HOG features and (ii) our eye detection method outperforms existing methods.

2 HOG Features

In this section, we briefly describe the HOG features. HOG features are inherited from the Scale Invariant Feature Transform proposed in [9]. They are derived based on a series of well-normalized local histograms of image gradient orientations in a dense grid [3]. The HOG feature extraction procedure is shown in Algorithm 1.

Algorithm 1. Overview of HOG Feature Extraction

Step1: Compute the horizontal and vertical gradient of image by convolving the image with a derivative mask.

Step2: Compute both norm and orientation of the gradient. Let G_h and G_v denote the horizontal and vertical gradient, respectively. The norm N_G and orientation O_G at the point (x, y) are given as follows:

$$N_G(x, y) = \sqrt{G_h(x, y)^2 + G_v(x, y)^2},$$

$$O_G(x, y) = arctan\frac{G_h(x,y)}{G_v(x,y)}.$$

Step3: Split the image into cells. Compute the histogram for each cell. Suppose the histogram is divided into K bins based on the orientation, the value of the i-th bin V_i for cell C is computed as follow:

$$V_i = \sum_{(x,y)\in C} \{N_G(x, y), O_G(x, y) \in Bin_i\}.$$

Step4: Normalize all histograms within a block of cell.

Step5: Concatenate all normalized histograms to form the HOG feature vector.

In our work, we use 1-D centered derivative $[-1, 0, 1]$ to compute the horizontal and vertical gradients. The size of cells is set to 4×4 pixels and the histogram is evenly divided into 6 bins over $0° - 180°$. Each block contains 3×3 cells and blocks are overlapped with each other by two-thirds in a sliding fashion. L2 normalization is used for block normalization scheme.

3 DHOG Features

In this section, we present a novel discriminating HOG (DHOG) features, which reside in low dimensional space and have significant discriminative power.

Let the extracted HOG feature vector introduced in Section 2 be $\mathcal{X} \in \mathbb{R}^N$, where N is the dimensionality of the HOG feature space. PCA is firstly applied to reduce the dimensionality of the original HOG feature from N to m, where $m < N$: $\mathcal{Y} = P^t\mathcal{X}$, where P contains the m eigenvectors of the covariance matrix of $\mathcal{X}$ corresponding to its m largest eigenvalues: $\lambda_1, \lambda_2, \cdots, \lambda_m$.

After PCA, the new feature vector $\mathcal{Y}$ resides in a lower dimensional space $\mathbb{R}^m$. In this $\mathbb{R}^m$ dimensional space, we implement the whitening transformation to sphere the covariance matrix of $\mathcal{Y}$. The whitening transformation is defined by the transformation matrix W: $W = \mathit{\Gamma}^{-1/2}P$, where $\mathit{\Gamma} = diag(\lambda_1, \lambda_2, \cdots, \lambda_m)$.

Next, we will define two scatter vectors and a criterion vector in order to derive the DHOG features. Let $W = \{W_1, W_2, \cdots, W_m\}$, where $W \in \mathbb{R}^{N \times m}$. Note that W contains m vectors. The idea of DHOG feature extraction is to choose a smaller set of vectors, from these m vectors, with the most discriminating capability. This smaller set of vectors will be the basis vectors for defining the DHOG feature. Toward that end, we first define the within-class scatter vector, $\alpha \in \mathbb{R}^m$, and the between-class scatter vector, $\beta \in \mathbb{R}^m$, as follows:

$$\alpha = P_1 \sum_{i=1}^{n_1} s(W^t x_i^{(1)} - W^t M_1) + P_2 \sum_{i=1}^{n_2} s(W^t x_i^{(2)} - W^t M_2) \tag{1}$$

and

$$\beta = P_1 s(W^t M_1 - W^t M) + P_2 s(W^t M_2 - W^t M) \tag{2}$$

where P_1 and P_2 are the prior probabilities, n_1 and n_2 are the number of samples, and $x_i^{(1)}$ and $x_i^{(2)}$ are the HOG features of the eye and the noneye samples, respectively. M_1, M_2, and M are the means of the eye class, the noneye class, and the grand mean in the original HOG feature space, respectively. The $s(\cdot)$ function defines the absolute value of the elements of the input vector. The significance of this new scatter vectors is that the within-class scatter vector, $\alpha \in \mathbb{R}^m$, measures the clustering capability of the vectors in W, and the between-class scatter vector, $\beta \in \mathbb{R}^m$, measures the separating capability of the vectors in W. In order to choose the most discriminating vectors from W to form a set of basis vectors to define DHOG features, we then define a new criterion vector $\gamma \in \mathbb{R}^m$, as follows:

$$\gamma = \beta./\alpha \tag{3}$$

where ./ is element-wise division. The value of the elements in γ indicates the discriminating power of their corresponding vectors in W: the larger the value is, the more discriminating power the corresponding vector in W possesses. Therefore, we choose the p vectors, $W_{i1}, W_{i2}, \cdots, W_{ip}$, in W corresponding to the p largest values in γ to form a basis $T = [W_{i1}, W_{i2}, \cdots, W_{ip}]$, where $T \in \mathbb{R}^{N \times p}$ and $p < m$. The DHOG features are thus defined as follows:

$$\mathcal{Z} = T^t \mathcal{X} \tag{4}$$

We name $T \in \mathbb{R}^{N \times p}$ as the DHOG basis vectors. The DHOG features thus resides in the feature space $\mathbb{R}^p$ and capture the most discriminating HOG information of the original data X.

Note that our DHOG extraction method is different from the commonly used discriminant analysis methods, such as Fisher Linear Discriminant (FLD) [5]. FLD seeks a set of basis vectors that maximizes the criterion $J = trace(S_w^{-1} S_b)$ [5], where S_w and S_b are the within-class and between-class scatter matrices. The criterion is maximized when the basis vectors are the eigenvectors of the matrix $S_w^{-1} S_b$ corresponding to its largest eigenvalues. FLD can find up to $L - 1$ basis vectors for the L-class pattern recognition problem. For a two-class eye detection problem, FLD is just able to derive only one feature, while our DHOG method is able to derive multiple features for achieving more reliable eye detection results.

4 Experiments

We evaluate the performance of DHOG features and our eye detection method on the Face Recognition Grand Challenge (FRGC) version 2 experiment 4, which contains both controlled and uncontrolled images [11]. Note that while the faces in the controlled images have good image resolution and illumination, the faces in the uncontrolled images have lower image resolution and large illumination variations. In addition, facial expression changes are in a wide range from open eyes to closed eyes, from without glasses to with various glasses, from black pupils to red and blue pupils, from white skin to black skin, and from long hair to wearing a hat. All these factors increase the difficulty of accurate eye-center detection. In our experiments, we do the test on the whole training data set of FRGC 2.0, which contains 12,776 images. So there are 25,552 eyes totally to be detected. In order to train a robust eye detector, 3,000 pairs of eyes and 12,000 non-eye patches are collected as training samples from different sources.

In Fig. 2 - Fig. 4, we compare the detection accuracy of DHOG features with the standard HOG features through different pixel errors. The HOG features after PCA (PHOG) are also included into the comparison to show the superior performance of DHOG. We don't list the result of the HOG features after FLD, since its performance is relatively lower in our experiments and thus is not a good criterion to evaluate the DHOG performance. A near neighbor classifier with three different distance metrics - L1 (city-block), L2 (Euclidean), and Cosine - are employed for classification. The size of the standard HOG features in our experiment is 1,296. The size of both PHOG and DHOG features is set to 80 for the best performance and fair comparison. The detection accuracy is measured as the Euclidean distance between the detected eye and the ground truth.

From Fig. 2 - Fig. 4, it is observed that no matter what kind of distance metric is applied, DHOG features outperform both of HOG and PHOG. In average, DHOG improves the detection accuracy of HOG by 1.39% and PHOG by 2.07%, respectively. If we consider the eye is detected correctly when the Euclidean distance between the detected eye and the ground truth is less than 5 pixels, DHOG reaches the best detection rate of 92.25% under COS metric, compared with 90.58% of HOG under L1 and 89.66% of PHOG under COS, respectively (see Table 1).

Table 1 lists the average detection pixel errors of all three methods in order to further show the performance improvement of DHOG over HOG and PHOG. Table 1 indicates that DHOG has smaller average detection pixel errors and higher detection rate than HOG and PHOG. The best result for each method under different distance metrics is highlighted. The DHOG+COS reaches the best, whose Euclidean distance error is about 3.16 pixels. We plot the distribution of eye detection pixel error for the best performance of each method in Fig. 5. It is observed that DHOG gives more detections in the range of one and five pixels from the ground truth than HOG and PHOG, which indicates the more accurate detection performance of DHOG.

We then compare our DHOG-based method with other eye detection methods. Although authors do not think the normalized errors is not a strict criterion to

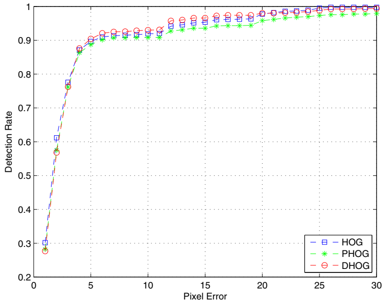


Fig. 2. Detection rate under L1 distance

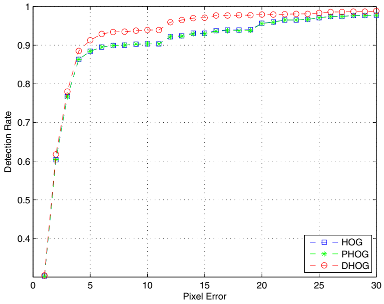


Fig. 3. Detection rate under L2 distance

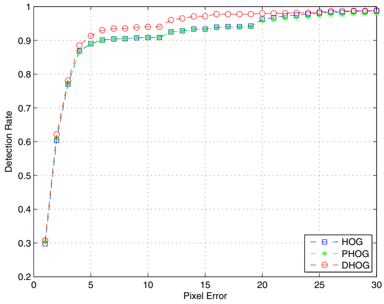


Fig. 4. Detection rate under COS distance

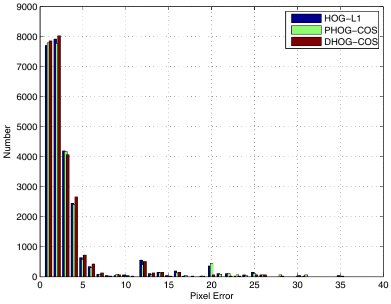


Fig. 5. Distribution of pixel errors

Table 1. Performance comparison among HOG, PHOG, and DHOG under different distance metrics (ED stands for the Euclidean distance)

Method	mean(x)	std(x)	mean(y)	std(y)	ED(mean)	Detection Rate
HOG+L1	2.50	2.68	2.24	4.12	3.53	90.58%
HOG+L2	2.46	2.63	2.33	4.20	3.83	89.16%
HOG+COS	2.45	2.60	2.45	4.49	3.79	89.66%
PHOG+L1	2.52	2.59	2.26	4.08	3.81	89.58%
PHOG+L2	2.45	2.60	2.36	4.26	3.84	89.16%
PHOG+COS	2.43	2.56	2.33	4.21	3.77	89.66%
DHOG+L1	2.57	2.65	1.92	3.24	3.40	91.24%
DHOG+L2	2.39	2.33	1.78	3.02	3.19	92.16%
DHOG+COS	2.38	2.32	1.76	2.91	3.16	92.25%

measure the performance of an eye detection method as explained in Section 1, it is still introduced here in order to make a fair comparison. The normalized error is the pixel error normalized by the binocular distance. Fig. 6 shows a typical comparison of our DHOG+COS method, which is reported with the best performance in our experiment, with the hybrid classifier of Jin in [6], who reported results on 3816 images of FERET database, and with the SVM based method of Campadelli in [1], who reported results on 862 images of FRGC 1.0 database. It is observed from Fig. 6 that although our experiments are performed in a much larger database (12,776 images) with more challenging compliancy (various illumination, poses, expressions, and obstructions), our eye detection method still outperform these two methods.

Finally, some examples of the detection result are listed in Fig. 7.

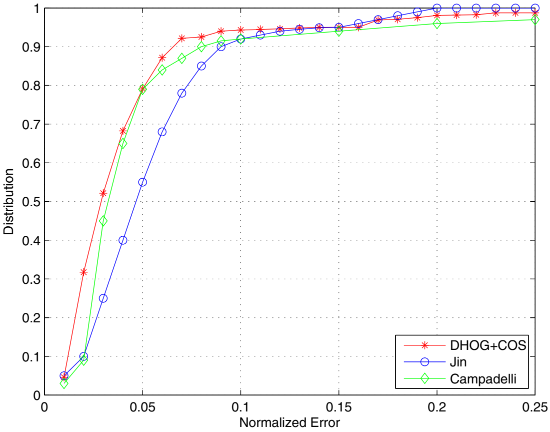


Fig. 6. Comparison of normalized detection error with different methods

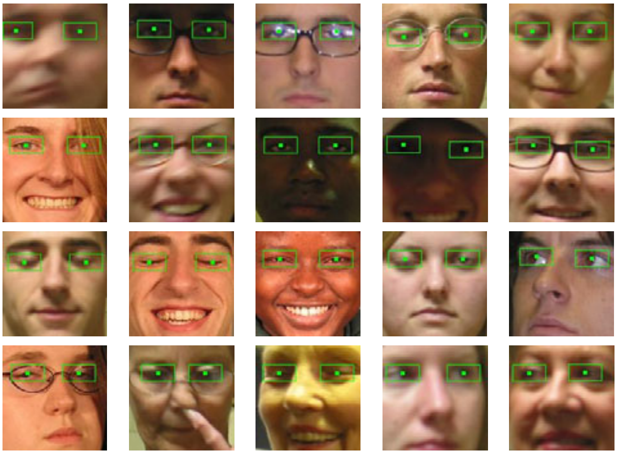


Fig. 7. Example of detected eyes

5 Conclusion

In this paper, we present a precise eye detection method using Discriminating HOG (DHOG) features. The DHOG features reside in a low dimensional space spanned by a set of DHOG basis vectors and have improved discriminating power over the standard HOG features. Experiments on FRGC database show that (i) DHOG features enhance the discriminating power of HOG features and (ii) our eye detection method outperforms the existing methods. Future work will focus on designing an automatic face recognition system using the DHOG features and eye detection method presented in this paper.

References

1. Campadelli, P., Lanzarotti, R., Lipori, G.: Precise eye localization through a general-to-specific model definition. In: British Machine Vision Conference (2006)
2. Chen, S., Liu, C.: Eye detection using color information and a new efficient svm. In: IEEE Int. Conf. on Biometrics: Theory, Applications and Systems (2010)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition, pp. 886–893 (2005)
4. Eckhardt, M., Fasel, I., Movellan, J.: Towards practical facial feature detection. Internatioanl Journal of Pattern Recognition and Artificial Intelligence 23(3), 379–400 (2009)
5. Fukunaga, K.: Introduction to statistical pattern recognition. Academic Press, London (1990)
6. Jin, L., Yuan, X., Satoh, S., Li, J., Xia, L.: A hybrid classifier for precise and robust eye detection. In: IEEE Int. Conf. on Pattern Recognition (2006)
7. Kroon, B., Maas, S., Boughorbel, S., Hanjalic, A.: Eye localization in low and standard definition content with application to face matching. Computer Vision and Image Understanding 113(4), 921–933 (2009)
8. Liu, C.: A Bayesian discriminating features method for face detection. IEEE Trans. Pattern Analysis and Machine Intelligence 25(6), 725–740 (2003)
9. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
10. Nguyen, M., Perez, J., Frade, F.: Facial feature detection with optimal pixel reduction svm. In: IEEE International Conference on Automatic Face and Gesture (2008)
11. Phillips, P., Flynn, P., Scruggs, T.: Overview of the face recognition grand challenge. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition (2005)
12. Phillips, P., Moon, H., Rizvi, S., Rauss, P.: The feret evaluation methodology for face recognition algorithms. IEEE Transaction on Pattern Analysis and Machine Intelligence 22(10), 1090–1104 (2000)
13. Wang, P., Green, M., Ji, Q., Wayman, J.: Automatic eye detection and its validation. In: IEEE International Conference on Computer Vision and Pattern Recognition (2005)
14. Wang, P., Ji, Q.: Multi-view face and eye detection using discriminant features. Computer Vision and Image Understanding 105(2), 99–111 (2007)
15. Zhu, Z., Ji, Q.: Robust real-time eye detection and tracking under variable lighting conditions and various face orientations. Computer Vision and Image Understanding 98(1), 124–154 (2005)

Detection of Retinal Vascular Bifurcations by Trainable V4-Like Filters

George Azzopardi and Nicolai Petkov

Johann Bernoulli Institute for Mathematics and Computer Science,
University of Groningen, The Netherlands
{g.azzopardi,n.petkov}@rug.nl

Abstract. The detection of vascular bifurcations in retinal fundus images is important for finding signs of various cardiovascular diseases. We propose a novel method to detect such bifurcations. Our method is implemented in trainable filters that mimic the properties of shape-selective neurons in area V4 of visual cortex. Such a filter is configured by combining given channels of a bank of Gabor filters in an AND-gate-like operation. Their selection is determined by the automatic analysis of a bifurcation feature that is specified by the user from a training image. Consequently, the filter responds to the same and similar bifurcations. With only 25 filters we achieved a correct detection rate of 98.52% at a precision rate of 95.19% on a set of 40 binary fundus images, containing more than 5000 bifurcations. In principle, all vascular bifurcations can be detected if a sufficient number of filters are configured and used.

Keywords: DRIVE, Gabor filters, retinal fundus, trainable filters, V4 neurons, vessel bifurcation.

1 Introduction

The vascular topographical geometry in the retina is known to conform to structural principles that are related to certain physical properties [14]. The analysis of the geometrical structure is very important as deviations from the optimal principles may indicate some cardiovascular diseases, such as hypertension [17] and atherosclerosis [4]; a comprehensive analysis is given in [12]. The identification of vascular bifurcations is one of the basic steps in this analysis.

More than 100 vascular bifurcations can be seen in a typical retinal fundus image. Their manual detection by a human observer is a tedious and time consuming process. The existing attempts to automate the detection of retinal vascular bifurcations can be categorized into two classes usually referred to as geometrical-feature based and model based approaches. The former involve extensive preprocessing such as segmentation and skeletonization followed by local pixel processing and branch point analysis. These techniques are known for their robustness in bifurcation localization [2,3,5,8]. On the other hand, model based approaches are usually more adaptive and have smaller computational complexity which makes them more appropriate for real-time applications [1,16].

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 451–459, 2011.

However, model based approaches are known to suffer from insufficient generalization ability as they are usually unable to model all the features of interest. Consequently, these methods may fail to detect some relevant features.

In this paper we propose trainable filters for the detection of vascular bifurcations in retinal fundus images. Our approach requires a single-step training process where an observer specifies a typical bifurcation by a point of interest in an image. The specified feature is then used to automatically configure a bifurcation detector by determining the properties of all line segments in the concerned feature and their mutual geometrical arrangement. This training procedure can be repeated as many times as required in order to configure a number of filters based on different specified features of interest. The filters can then be applied on retinal fundus images to detect the features that are similar to the patterns that were used to configure the filters.

The rest of the paper is organized as follows: In Section 2 we present our method and demonstrate how it can be used to detect retinal vascular bifurcations. In Section 3, we apply the proposed nonlinear filters on retinal fundus images from the DRIVE dataset [15]. Section 4 contains a discussion and conclusions.

2 Proposed Method

2.1 Overview

Fig. 1a shows a bifurcation encircled in a binarized retinal fundus image from the DRIVE dataset [15]. Such a feature, which is shown enlarged in Fig. 1b, is used to automatically configure a detector that will respond to the same and similar patterns.

Each of the three ellipses shown in Fig. 1b represents the support or receptive field (RF) of a sub-unit that detects a line of a given orientation and width, while the central circle represents the RF of a group of such sub-units. The response of the proposed bifurcation detector is computed by combining the responses of the concerned sub-units by multiplication. The preferred orientations of the sub-units and the mutual spatial arrangement of their RFs are determined by the local pattern used for the configuration of the concerned filter. Consequently,

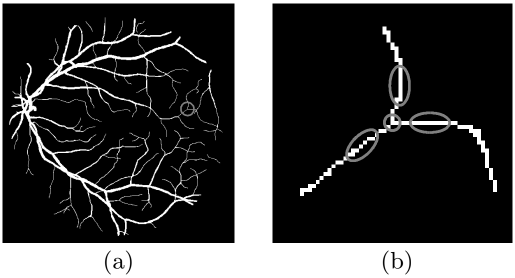


Fig. 1. (a) The circle indicates a bifurcation that is selected by a user. (b) Enlargement of the selected feature. The ellipses represent the support of line detectors that are identified as relevant for the concerned feature.

that filter is selective for the presented local combination of lines of specific orientations and widths.

Such a design is inspired by electrophysiological evidence that some neurons in area V4 of visual cortex are selective for moderately complex stimuli, such as curvatures, that receive inputs from a group of orientation-selective cells in areas V1 and V2 [9,10,11]. Moreover, there is psychophysical evidence [6] that curve contour parts are likely detected by an AND-gate-like operation that combines the responses of afferent orientation-selective sub-units by multiplication. An AND-gate-like model produces a response only when all its afferent sub-units are stimulated; i.e. all constituent parts of a stimulus are present.

In the next sub-sections, we explain the automatic configuration process of a bifurcation detector. The configuration process determines which responses of which Gabor filters in which locations need to be multiplied in order to obtain the output of the filter.

2.2 Orientation-Selective Sub-units Based on Gabor Filters

The input to the orientation-selective sub-units mentioned above is provided by two-dimensional (2D) Gabor filters, which are established models of V1/V2 cells. We denote by $g_{\lambda,\theta}(x, y)$ the half-wave rectified response of a Gabor filter of preferred wavelength λ and orientation θ to a given input image. Such a filter has also other parameters, namely spatial aspect ratio, bandwidth and phase offset, that we skip here for brevity. We set their values as proposed in [13].

Since we work in a multiscale setting, we re-normalize all Gabor functions that we use in such a way that all positive values of such a function sum up to 1 while all negative values sum up to -1. We use symmetric Gabor functions as they respond to line structures and we are interested to detect the presence of vessels in retinal fundus images.

We use a bank of Gabor filters with 5 wavelengths ($\varLambda = \{4, 4\sqrt{2}, 8, 8\sqrt{2}, 16\}$) and 8 equidistant orientations ($\varTheta = \{0, \frac{\pi}{8}, \ldots, \frac{7\pi}{8}\}$) that we apply on images of size 565 × 584. In such images, the blood vessels have widths of 1 to 7 pixels. Fig.2a illustrates the maximum value superposition of the thresholded responses of the concerned bank of Gabor filters obtained for the bifurcation image shown in Fig.1b. All responses are thresholded at a given fraction $t_1 = 0.2$ of the maximum response of $g_{\lambda,\theta}(x, y)$ across all combinations of values (λ, θ) used and all positions (x, y) in the image.

2.3 Sub-unit Parameters

A sub-unit uses as inputs the responses of a certain Gabor filter characterized by the parameter values (λ, θ) around a certain position (ρ, ϕ) with respect to the center of the filter. A sub-unit is thus characterized by four parameters: $(\lambda, \theta, \rho, \phi)$. The values of such parameters for a sub-unit are obtained as follows.

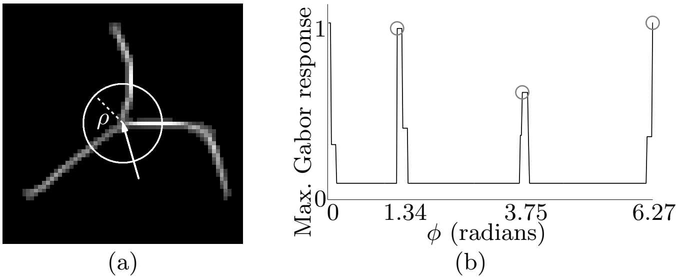
We consider the responses of the bank of Gabor filters along a circle of a given radius ρ around the selected point of interest (Fig.2). In each position along that circle, we take the maximum of all responses across the possible values of (λ, θ).

If this value is greater than the corresponding values for the neighboring positions along an arc of angle $\frac{\pi}{8}$ the concerned position is chosen as a center of the RF of a sub-unit. Its coordinates (ρ, ϕ) are determined with respect to the center of the filter. The pair of values (λ, θ) for which the concerned local maximum is reached are the preferred wavelength and orientation of the sub-unit.

In our experiments, we configure bifurcation detectors using multiple values of the parameter ρ. For non-zero values of ρ we determine a group of sub-units with the method mentioned above. For $\rho = 0$, we consider the responses of the bank of Gabor filters used at the specified point of interest. For such a location, we consider all combinations of (λ, θ) for which the corresponding responses $g_{\lambda,\theta}(x, y)$ are greater than a fraction $t_2 = 0.75$ of the maximum of $g_{\lambda,\theta}(x, y)$ across the different combinations of values (λ, θ) used. For each value θ that satisfies such a condition, we consider a single value of λ, the one for which $g_{\lambda,\theta}(x, y)$ is the maximum of all responses across all values of λ. At this central location, multiple sub-units can thus be defined and their RFs are centered at the same position with polar coordinates $(\rho = 0, \phi = 0)$.

We denote the set of parameter value combinations, which fulfill the above conditions, by $S_f = \{(\lambda, \theta, \rho, \phi)\}$. The subscript f stands for the local pattern around the selected point of interest. Every tuple in the set S_f specifies the parameters of a sub-unit.

For the point of interest shown in Fig.2a and two given values of the radius ρ ($\{0, 10\}$), the selection method described above results in five sub-units with parameter values specified by the tuples in the following set; $S_f = \{(\lambda = 4, \theta = 0, \rho = 0, \phi = 0), (\lambda = 4, \theta = \frac{\pi}{2}, \rho = 0, \phi = 0), (\lambda = 4, \theta = 0, \rho = 10, \phi = 1.34), (\lambda = 4, \theta = \frac{3\pi}{4}, \rho = 10, \phi = 3.75), (\lambda = 4, \theta = \frac{\pi}{2}, \rho = 10, \phi = 6.27)\}$. The last tuple in that list, $(\lambda = 4, \theta = \frac{\pi}{2}, \rho = 10, \phi = 6.27)$, for instance, describes a sub-unit that collects its inputs from the responses of a Gabor filter with $\lambda = 4$ and $\theta = \frac{\pi}{2}$, i.e. a Gabor filter that strongly responds to horizontal lines ($\theta = \frac{\pi}{2}$)



(a) (b)

Fig. 2. (a) The gray-level intensity of every pixel is the maximum value superposition of the thresholded responses from a bank of Gabor filters at that position; $\max_{\lambda\in\Lambda,\theta\in\Theta} |g_{\lambda,\theta}(x, y)|_{t_1}$. The arrow indicates the location of the point of interest selected by a user, while the bright circle of a given radius ρ indicates the considered locations. (b) Values of the maximum value superposition of Gabor filter responses along the concerned circle of radius $\rho = 10$ around the point of interest. The marked local maxima are caused by the three blood vessels.

of width of ($\frac{\lambda}{2}$ =) 2 pixels, around a position of (ρ =) 10 pixels to the right ($\phi = 6.27$) of the center of the filter. This selection is the result of the presence of a horizontal vessel to the right of the center of the feature that is used for the configuration of the filter.

2.4 Sub-unit Response

We denote by $s_{\lambda,\theta,\rho,\phi}(x, y)$ the response of a sub-unit, which we compute as follows. We consider the responses $g_{\lambda,\theta}(x, y)$ of a Gabor filter with preferred wavelength λ and orientation θ around position (ρ, ϕ) with respect to the center of the filter. We weight these responses by a 2D Gaussian function with a standard deviation that is a linear function of parameter ρ. We define the output of the sub-unit as the maximum value of all the weighted responses of the concerned Gabor filter. This result is shifted by ρ in the direction opposite to ϕ.

Fig.3 illustrates the computation of the responses of three sub-units. Each of the three bright blobs shown is an intensity map of a 2D Gaussian function mentioned above. The three ellipses illustrate the orientations and wavelengths of the corresponding Gabor filters. The responses $g_{\lambda,\theta}(x, y)$ of such a filter are weighted by the respective 2D Gaussian function and the maximum result is shifted by the corresponding vector.

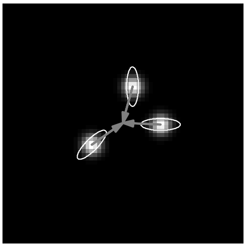


Fig. 3. Computation of sub-unit responses. The three bright blobs are intensity maps for 2D Gaussian functions that model the corresponding sub-unit RFs. The three ellipses illustrate the orientations and wavelengths of the corresponding Gabor filters. A sub-unit response is computed as the maximum value of the weighted responses of such a Gabor filter with the respective 2D Gaussian function. The result is shifted by the corresponding vector.

2.5 Filter Response

We define a nonlinear filter with output r_{S_f} as the geometric mean of all quantities $s_{\lambda,\theta,\rho,\phi}(x, y)$ that belong to the specific selection determined by S_f:

$$r_{S_f}(x, y) = \left| \left(\prod_{(\lambda,\theta,\rho,\phi)\in S_f} s_{\lambda,\theta,\rho,\phi}(x, y) \right)^{\frac{1}{|S_f|}} \right|_{t_3} \tag{1}$$

where $|.|_{t_3}$ stands for thresholding the response at a fraction t_3 of its maximum.

Rotation invariance is achieved by manipulating the set of parameter values in S_f, rather than by computing them from the responses to a rotated version of the original pattern. Using the set S_f that defines the concerned filter, we can form a new set $\Re_\psi(S_f) = \{(\lambda, \theta + \psi, \rho, \phi + \psi) \mid (\lambda, \theta, \rho, \phi) \in S_f\}$. The rotation invariant response is then defined as $\widehat{r}_{S_f}(x, y) = \max_\psi(r_{\Re_\psi(S_f)}(x, y))$.

3 Experimental Results

We use the bifurcation illustrated in Fig.1 to configure a filter denoted by S_{f_1} ($\rho \in \{0, 4, 10\}$, $t_1 = 0.2$ and $t_2 = 0.75$). Fig.4(a-b) show the result (for $t_3 = 0.25$) of the application of filter S_{f_1} to the binary retinal fundus image shown in Fig.1a. The encircled regions are centered on the local maxima of the filter response and if two such regions overlap by 75%, only the one with the stronger response is shown. Besides the original bifurcation, the filter successfully detects 5 other bifurcations with similar vessel orientations.

If the same filter is applied in a rotation invariant mode, a total of 38 similar features are detected, Fig.4(c-d). This illustrates the strong generalization capability of this approach because 35.51% (38 out of 107) of the features of interest are detected by a single filter. Notable is the fact that this is achieved at a precision rate of 100%, as the filter does not give any false positive responses. The threshold parameter t_3 can be used to tune the degree of generalization.

As to the remaining features that are not detected by this filter, we proceed as follows: we take one of these features that we denote by f_2 (Fig. 5) and train a second nonlinear filter, S_{f_2}, using it. With this second filter we detect 46 features of interest of which 20 coincide with features detected by filter S_{f_1} and 26 are newly detected features. Merging the responses of the two filters results in the detection of 64 distinct features. We continue adding filters that are configured using features that have not been detected by the previously trained filters. A set of 10 filters that correspond to the features shown in Fig.5 proves sufficient to detect all 107 features of interest in the concerned image. A fixed response

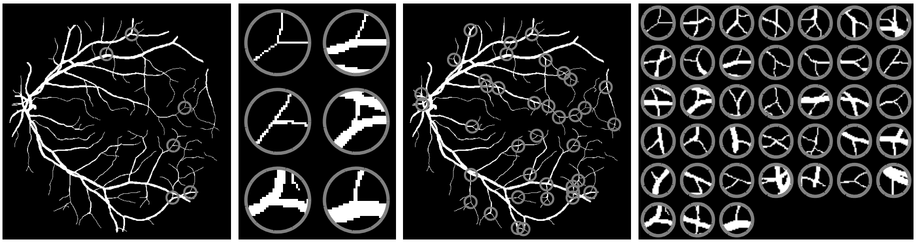


Fig. 4. (a) Result of applying the filter S_{f_1} in rotation non-invariant mode and (b) enlargements of the detected features given in descending order (left-to-right, top-to-bottom) of the filter response. (c) Result of applying the filter in a rotation invariant mode and (d) enlargements of the detected features.

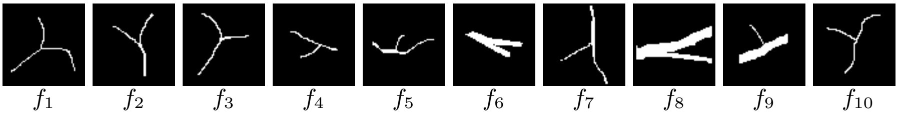


Fig. 5. A set of 10 bifurcations extracted from the image in Fig. 1a, used to configure 10 filters

threshold of $t_3 = 0.25$ is applied for all filters. An important aspect of this result is that a recall rate of 100% is achieved at a precision rate of 100% [1].

We apply these 10 filters on a larger dataset (DRIVE) of 40 binary retinal fundus images[2]. The ground truth of correct bifurcations was defined by the authors of this paper. For this larger dataset we achieve a recall rate R of 97.3% and a precision rate P of 94.71%. We carried out further experiments by configuring up to 40 filters and varying the threshold parameter t_3 between 0.2 and 0.3. We achieve optimal results for 25 filters and show them together with the results for 10 filters in Fig. 6. With 25 filters, the harmonic mean $(2PR/(P+R))$ of the precison and recall reaches maximum at a recall rate of 98.52% and a precision rate of 95.19% for $t_3 = 0.28$.

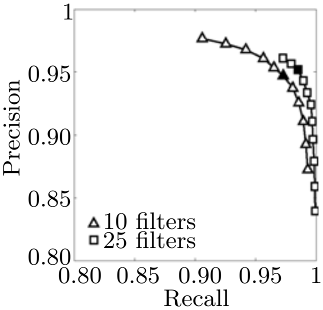


Fig. 6. Precision-recall plots obtained with 10 and 25 filters. For each plot the threshold parameter t_3 is varied between 0.2 and 0.3. The precision rate increases and the recall rate decreases with an increasing value of t_3. The harmonic mean of precision and recall reaches a maximum at $R = 0.9852$ and $P = 0.9519$ for 25 filters and at $R = 0.973$ and $P = 0.9471$ for 10 filters. These points are marked by a filled-in square and triangle, respectively.

4 Discussion and Conclusion

We propose a novel approach for the automation of vascular bifurcation detection in retinal fundus images. Our proposed method is implemented in filters that simulate the properties of shape-selective V4 neurons in visual cortex.

[1] Recall rate is the percentage of true bifurcations that are successfully detected. Precision rate is the percentage of correct bifurcations from all detected features.

[2] Named in DRIVE 01_manual1.gif, 02_manual1.gif, . . . , 40_manual1.gif

The proposed V4-like filters are trainable, in that the structure of the filter is determined by a feature that is specified by a user. The way this is achieved is not by template matching, but rather by the extraction of information about the dominant orientations in the concerned feature and their mutual spatial arrangement. While such a filter reacts most strongly to the feature that was used to configure it, the filter also reacts to features which differ in the orientations of the involved line segments to a certain extent. The degree of generalization can be tuned by proper selection of the filter parameters. The automatic configuration of the proposed filters gives an edge to our approach over model based approaches regarding generalization ability.

Although one can find methods for local image feature analysis by combining filter responses at different scales (e.g. SIFT features [7]), to the best of our knowledge, the proposed approach is the first one which combines the responses of orientation-selective filters with their main area of support outside the point of interest.

In our experiments, we use a set of 40 binary retinal images provided as ground truth in the DRIVE dataset [15]. In total, these images contain 5118 vessel bifurcations. We achieved a recall rate of 98.52% and a precision rate of 95.19% with the application of only 25 filters. The precision rate can be improved by performing additional analysis of the features that are detected by the filters. In [2] a recall rate of 95.82% was reported on a small dataset of five retinal images.

In principle, all vessel bifurcations can be detected if a sufficient number of filters are configured and used. The recall rate of 98.52% that we achieve means that on average only one to two out of 100 bifurcations are missed in a typical image. This is sufficient to the needs of the medical application at hand. We conclude that the proposed trainable filters are an effective means to automatically detect bifurcations in retinal vascular images.

References

1. Ali, C., Hong, S., Turner, J., Tanenbaum, H., Roysam, B.: Rapid automated tracing and feature extraction from retinal fundus images using direct exploratory algorithms. IEEE Transactions on Information Technology in Biomedicine 3, 125–138 (1999)
2. Bhuiyan, A., Nath, B., Chua, J., Ramamohanarao, K.: Automatic detection of vascular bifurcations and crossovers from color retinal fundus images. In: Third International IEEE Conference on Signal-Image Technologies and Internet-Based System (SITIS), pp. 711–718 (2007)
3. Chanwimaluang, T., Guoliang, F.: An efficient blood vessel detection algorithm for retinal images using local entropy thresholding. In: Proceedings of the 2003 IEEE International Symposium on Circuits and Systems (Cat. No.03CH37430) 5
4. Chapman, N., Dell'omo, G., Sartini, M., Witt, N., Hughes, A., Thom, S., Pedrinelli, R.: Peripheral vascular disease is associated with abnormal arteriolar diameter relationships at bifurcations in the human retina. Clinical Science 103 (2002)
5. Eunhwa, J., Kyungho, H.: Automatic retinal vasculature structure tracing and vascular landmark extraction from human eye image. In: International Conference on Hybrid Information Technology, vol. 7 (2006)

6. Gheorghiu, E., Kingdom, F.: Multiplication in curvature processing. Journal of Vision 9 (2009)
7. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
8. Martinez-Perez, M., Hughes, A., Stanton, A., Thom, S., Chapman, N., Bharath, A., Parker, K.: Retinal vascular tree morphology: A semi-automatic quantification. IEEE Transactions on Biomedical Engineering 49, 912–917 (2002)
9. Pasupathy, A., Connor, C.: Responses to contour features in macaque area v4. Journal of Neurophysiology 82, 2490–2502 (1999)
10. Pasupathy, A., Connor, C.: Shape representation in area v4: Position-specific tuning for boundary conformation. Journal of Neurophysiology 86, 2505–2519 (2001)
11. Pasupathy, A., Connor, C.: Population coding of shape in area v4. Nature Neuroscience 5, 1332–1338 (2002)
12. Patton, N., Aslam, T., MacGillivray, T., Deary, I., Dhillon, B., Eikelboom, R., Yogesan, K., Constable, I.: Retinal image analysis: Concepts, applications and potential. Progress in Retinal and Eye Research 25, 99–127 (2006)
13. Petkov, N.: Biologically motivated computationally intensive approaches to image pattern-recognition. Future Generation Computer Systems 11, 451–465 (1995)
14. Sherman, T.: On connecting large vessels to small - the meaning of murray law. Journal of General Physiology 78, 431–453 (1981)
15. Staal, J., Abramoff, M., Niemeijer, M., Viergever, M., van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. IEEE Transactions on Medical Imaging 23, 501–509 (2004)
16. Tsai, C., Stewart, C., Tanenbaum, H., Roysam, B.: Model-based method for improving the accuracy and repeatability of estimating vascular bifurcations and crossovers from retinal fundus images. IEEE Transactions on Information Technology in Biomedicine 8, 122–130 (2004)
17. Tso, M., Jampol, L.: Path-physiology of hypertensive retinopathy. Opthalmology 89 (1982)

A Method for Identification and Visualization of Histological Image Structures Relevant to the Cancer Patient Conditions

Vassili Kovalev[1], Alexander Dmitruk[1], Ihar Safonau[1],
Mikhail Frydman[2], and Sviatlana Shelkovich[3]

[1] Biomedical Image Analysis Department, United Institute of Informatics Problems,
Surganova St., 6, 220012 Minsk, Belarus
[2] Department of Morbid Anatomy, Minsk City Hospital for Oncology,
Nezavisimosti av. 64-3, 220013 Minsk, Belarus
[3] Oncology Department, Belarusian Medical Academy of Post-Graduate Education,
Brovki St., 3-B3, 220013, Minsk, Belarus

Abstract. A method is suggested for identification and visualization of histology image structures relevant to the key characteristics of the state of cancer patients. The method is based on a multi-step procedure which includes calculating image descriptors, extracting their principal components, correlating them to known object properties and mapping disclosed regularities all the way back up to the corresponding image structures they found to be linked with. Image descriptors employed are extended 4D color co-occurrence matrices counting the occurrence of all possible pixel triplets located at the vertices of equilateral triangles of different size. The method is demonstrated on a sample of 952 histology images taken from 68 women with clinically confirmed diagnosis of ovarian cancer. As a result, a number of associations between the patients' conditions and morphological image structures were found including both easily explainable and the ones whose biological substrate remains obscured.

1 Introduction

It is well-known that visual examination of histological images taken from tissue samples remains a gold standard in definitive diagnosis, staging and treatment of a number of cancer types [1], [2]. However, the histological image analysis problem has not been adequately explored and remains underdeveloped comparing to other branches of recent image analysis methods [3], [4], [5]. This is mostly because the histological image data stay apart from the main body of biomedical images by their remarkable morphological complexity [3], [4], [6]. These holds true for majority of conventional methods and worsened even further by new emerging techniques of preprocessing of tissue probes and advanced imaging technologies such as the whole slide scanning producing hyper-large images [7].

Motivation. The motivation of this work stems from a biomedical problem of discovering implicit links between the morphological structure of histological

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 460–468, 2011.

images and features describing the state of cancer patients. In particular, we are interesting in attributing certain conditions of the ovarian cancer patients to morphological structures observed in routine diagnostic tissue samples as well as in the probes immuno-histochemically processed for highlighting tissue lympho- and angio-genesis.

Ovarian cancer is a devastating disease which is known as one of the major causes of female gynaecological death worldwide [8]. In Western countries about 1-2% of all women develop epithelial ovarian cancer at some time during their lives. The problem is also that the most patients refer to a hospital too late being already in an advanced stages of disease. This is because the first indication of the ovarian cancer is not a pain but simply swelling of the abdomen which can be easily missed. As a result, the five-year survival rates remain as low as 20% [8], [9]. This work is part of a larger project aimed at studying the malignant tumor angiogenesis in ovary [10]. Angiogenesis, the development of new blood vessels from the existing vasculature, is an important factor of solid tumor growth and metastasis [9]. Without angiogenesis the tumor expansion is naturally limited by 1-2 mm only because in order to grow the tumor needs to be supplied by oxygen and nutrition and waste removals outside [11]. Recently, there is a hope for cancer treatment by inhibiting angiogenesis processes. Thus, disclosing links between the tumor structure, its growth characteristics and patient conditions is of paramount importance for oncology [8], [9].

The technical problem. In a typical setup there is a patient database available which contains both image data of different modalities as well as non-visual patient characteristics such as general social data, clinical observations, results of laboratory tests, history of personal and family diseases, etc. Then technically the problem is posed as finding statistically significant associations between the morphological image structures presented in form of suitable quantitative features and database variables containing the patient records. Such correlations can be found in a straightforward manner using, for example, conventional approach of feature extraction followed by a multivariate statistical analysis for identifying significant links between these two. However, this is only possible with *a priori* research hypothesis in hands which presumes certain connections between the specific, pre-defined image structures and some patient characteristics. Being developed, implemented, and successfully applied to the input data, this approach leaves researcher with only particular results and image structures that have been extracted and examined. For instance, our preliminary study exploiting this approach was attempting to attribute tumor vessel development visualized with the help of D2-40 marker to patients' conditions. To this end, the vessel network was segmented, characterized by five quantitative features, and correlated to the patient state. However, despite certain time and other resources were spent, it gave very particular and rather modest results [10].

Thus, in this context it is worth to consider an alternative, exploratory approach which is aimed at detecting the whole bunch of objectively existing correlations between the histology image structures and patient state first and separate investigating their novelty together with the underlying biomedical substrate afterwards. Such an approach may conditionally be categorized into the image mining research area. In much the same way as data mining, the image mining can be understood as the process of extracting hidden patterns from images [12], [13]. More specifically, image mining deals with extraction of implicit knowledge, image data relationship or other patterns not explicitly stored in the image database (e.g., [14], [15], [16], [17]). Given that the histological image analysis is the task that difficult to automate due to its structural sophistication, it appears promising to examine the wide-cut image mining techniques for discovering the links we are interested in.

Basic requirements. In order to produce the desired result, a method of identification and visualization of histological image structures which correlate with the cancer patient conditions should fulfill the following major requirements.

(a) The image descriptors should be powerful and flexible enough to capture a broad range of morphological image properties and be capable of both color and grayscale images.
(b) The quantitative features which are derived from descriptors and correlated to patient state records should allow mapping selected correlations back to original images for isolating and visualizing underlying morphological structures.
(c) The number of features used for describing the image content should be limited by a few dozen to satisfy the well-known statistical requirement (i.e., kept less than the number of patients) what avoids correlation purely by chance.

In this paper, we introduce a method for discovering important histology image structures of cancer tissue that fulfill these requirements. We demonstrate its abilities on a database of 68 ovarian cancer patients.

2 Materials and Methods

Image Data. A database containing patient records and histological tissue images of 68 ovarian cancer patients (women, mean age 59.8 years, STD=11.2) was used with this study. The image data part consisted of 952 color images of 2048×1536 pixels in size which were acquired under ×200 magnification using recent Leica DMD108 microscope. They included 272 routine hematoxylin-eosin stained diagnostic images (4 images for each patient) and 680 images of tissue probes (10 per patient) immuno-histochemically processed with D2-40 marker highlighting lymphogenesis. Examples are provided in Fig. 1. Patients' state records included about 80 characteristics such as the international TNM cancer staging, medical history, tumor dissemination, surgery and chemotherapy details, current value of alive-died flag and some other.

The Method. Due to the characteristic textural appearance, color texture descriptors are the most common type of features used in histology image analysis

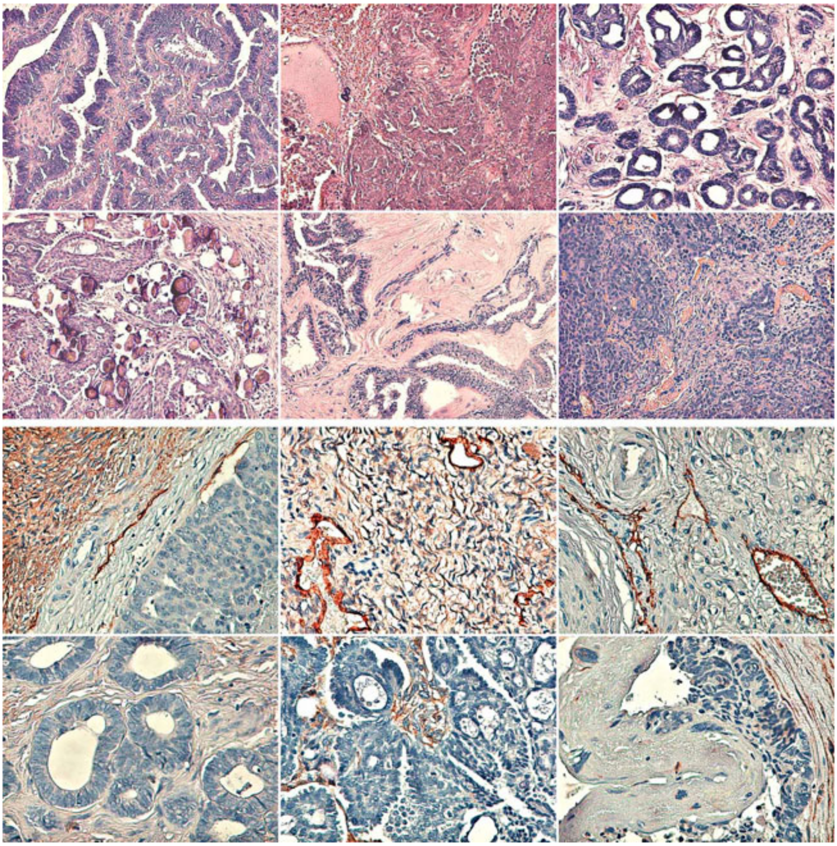


Fig. 1. Examples of original histological images of tissue routinely stained with hematoxylin-eosin (top two rows) and D2-40 endothelial marker (bottom two rows)

when describing the image as a whole. Among them are the color co-occurrence matrices introduced independently in [18] and [19] under the color correlogram term first [18] and as co-occurrence matrices a year later [19]. There are also several allied approaches for describing spatial image structure such as simultaneous autoregressive models [15] and some other. Here we continuing to exploit the co-occurrence approach. However, taking into account the first requirement given in the introduction, we developing the co-occurrence approach further and using extended 4D matrices. Namely, we considering triplets of pixels located at the vertices of equilateral triangles instead of conventional pixel pairs. Note that such an extension is not just mechanical addition of one more dimension to co-occurrence matrix array as this might appear at the first sight. The consequences are by far deeper and they related to the problem of discriminating different sorts of textures with the help of first (pixel intensities/colors alone), second (gradients), and higher order spatial statistics. This problem was thoroughly studied by Bela Julesz (e.g., [20]). More recently, this line of research on visual texture perception is studied with the help of fMRI brain scanning. In particular, it was experimentally proven [21] that patterns of brain activity are significantly different when observing textures with low and high order spatial correlations. Note that one should not mistake high order statistics in the *spatial* and in the *intensity* [22] domains.

Let $F_G = \{I(x,y)\} = \{I(i)\} = \{I(j)\} = \{I(k)\}$ be a gray-scale image of $M \times N$ pixels in size. Let suppose all the image pixels are indexed with the help of indices i, j, and k, where $i = \overline{1, MN}$, $j = \overline{1, MN}$ and $k = \overline{1, MN}$ and their intensity levels are $I(i)$, $I(j)$, and $I(k)$ respectively. The indices are naturally defined by pixel coordinates as $i = (x_i, y_i)$, $j = (x_j, y_j)$ and $k = (x_k, y_k)$. Then the 4D gray-scale intensity co-occurrence matrix of $IIID$ type defined on the triplets of pixels (i, j, k) which are located at the vertices of equilateral triangles with the side of d pixels can be defined as follows:

$$W_{IIID} = \| I(i), I(j), I(k), d \| ,$$

$$d(i,j) = d(i,k) = d(j,k), d \in D,$$

$$i < j, i < k,$$

$$\forall i : y_j \geq y_i, y_k < y_i.$$

Note that the last two lines of the above equation formalize the requirement of enumeration of all possible triangles with no repetition. The equation describes algorithm of covering the whole image by equilateral triangles. As it can be inferred, the procedure consists of subsequent placing the basic (seed) triangle vertex on the image position i so that the second vertex j falls into the same row for d pixels ahead with the vertex k pointing down. This gives the first, initial position with the seed vertex fixed at i whereas the rest ones are obtained by rotating the triangle around i clockwise so that its third, i.e. k-th vertex neither cross nor elevates over the current image row.

In case the image colors should be considered, the color space is suitably reduced first and corresponding color co-occurrence matrix of $CCCD$ type can be defined exactly in the same manner using the image color indices $C(i)$, $C(j)$, and $C(k)$ instead of intensity levels.

Once the co-occurrence matrices are calculated, the very common strategy is to calculate Haralick's features next and to use them for image characterization, clustering, etc. However, this traditional procedure may not be followed here at least because Haralick's features cannot be mapped back to the original images as the second introductory condition requires. On the contrary, the matrix elements themselves may be mapped back [23] but there are too many of them to satisfy the final, third condition. The solution is to apply PCA method for extracting a limited number of uncorrelated features from matrices.

Thus, the method supposes calculating 4D co-occurrence matrices, extracting principal components, correlating them to patients' state, selecting significant ones, projecting selected components back to co-occurrence matrix elements, and finally using them for visualizing the image structures we are looking for.

Note that since principal components are uncorrelated, there is no need to apply complicated and somewhat risky multivariate statistical analysis methods. Searching for significant links can be done by straightforward univariate correlations or with the help of Student's t-test according to the feature type.

3 Results

Original RGB images were converted into the *Lab* space with Euclidean color dissimilarity metrics and the number of colors was reduced down to 24 bins using the median cut algorithm preserving most important colors. Thus, the 3D color co-occurrence sub-matrices CCC with a fixed inter-pixel distance d contain $24^3 = 13824$ cells. Given that elements above leading diagonal are zeros, the number of effective matrix elements was $N_E = 2600$. Equilateral triangles with side lengths $D = \{1, 3, 5\}$ were considered so that the total number of elements of completed $CCCD$ matrices was 7800. Cumulative $CCCD$ matrices computed over all the images of each patient were vectorized constituting an input PCA data table with 68 rows and 7800 columns. PCA resulted in extracting 27 principal components (PCs) in case of matrices of routine images and 38 PCs in case of D2-40 images under condition of covering 95.0% of variance. The first components cover 55.7% and 26.5% of variances respectively. These results suggest that structural variability of D2-40 images is substantially higher compared to routine ones. Being correlated with patients' data, 27 PCs of routine images have produced a total of 43 events of correlation significant at $p < 0.01$. Same procedure being applied to 38 PCs derived from descriptors of D2-40 images with highlighted lymphatic vessels resulted in detecting 47 significant links between these features and patient state records.

Detailed investigation of significant correlation has revealed that some of them were easily deductible from existing knowledge whereas other are suggestive for novelty and certainly interesting from both scientific and practical points of view. For instance, in case of routine images the significant links between PCs and the following patient data appears to be very promising: development of distant metastases ($p < 0.001$), the degree of cancer tissue differentiation ($p < 0.007$), the number of miscarriages ($p < 0.0001$), and the number of chemotherapy trials ($p < 0.000002, r = -0.543$) (see visualization of related image structures on the top row of Fig. 2). The negative correlation of the length of borders highlighted in the figure with the number of trials may be explained by the fact that more spacious tumor structure is typical for relatively "young" tumors which are chemically treated first compared to "old" ones which removed immediately. Images of tissue processed by D2-40 endothelial marker have demonstrated similar behavior disclosing a number of interesting links. The bottom row of Fig. 2 demonstrates one of them which displays stromal structures (automatically extracted and visualized on the bottom-right picture) affected by proteins of endothelial cells. The fraction of these structures strongly correlates with tumor differentiation rate ($p < 0.009$), patient survival time ($p < 0.010$) and presence of a relapse ($p = 0.017$).

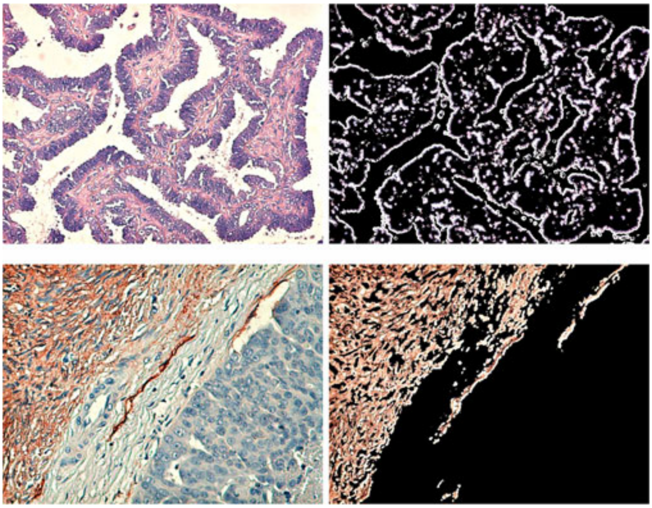


Fig. 2. Examples of original images (left column) and their key structures visualized (right column) for routine (top row) and D2-40 (bottom row) images

Finally, the abilities of *IIID* co-occurrence matrices computed using grayscale version of the images were also assessed. Despite some promising correlations were found, an ambiguity was revealed. In particular, when certain *IIID* matrix element was mapped back to the grayscale images, it highlights structures of biologically different sorts. This is because two or more substantially different image colors were converted down to one single gray level.

4 Conclusions

The results reported with this study allow to draw the following conclusions.

1. The method presented in this paper may be considered as a promising tool capable of an automatic identification and visualization of histological image structures relevant to the cancer patient conditions.
2. Since there is no intrinsic mechanism for semantic assessing the resultant links detected by the method, an expert-based evaluation of the novelty and biological substrate of the result is necessary.
3. The future work should include development of an automatic procedure for selecting the set of matrix elements to be mapped back to original images once the interesting principal component is identified.

Acknowledgments. This work was funded by the ISTC project B–1682.

References

1. Schwab, M.: Encyclopedia of Cancer, 2nd edn., 4 volumes, 3235p. Springer, Heidelberg (2009)
2. Hayat, M.A.: Methods of Cancer Diagnosis, Therapy and Prognosis, 6 volumes. Springer, Heidelberg (2009/2010)
3. Wootton, R., Springall, D., Polak, J.: Image Analysis in Histology: Conventional and Confocal Microscopy, 425p. Cambridge University Press, Cambridge (1995)
4. Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B.: Histopathological image analysis: A review. IEEE Reviews in Biomedical Engineering (1), 147–171 (2009)
5. Sertel, O., Kong, J., Catalyurek, U., Lozanski, G., Saltz, J., Gurcan, M.: Histopathological image analysis using model-based intermediate representations and color texture: Follicular lymphoma grading. Journal of Signal Processing Systems 55(1), 169–183 (2009)
6. Yu, F., Ip, H.: Semantic content analysis and annotation of histological images. Computers in Biology and Medicine 38(6), 635–649 (2008)
7. Rojo, M.G., Garcia, G.B., Mateos, C.P., Garcia, J.G., Vicente, M.C.: Critical comparison of 31 commercially available digital slide systems in pathology. International Journal of Surgical Pathology 14(4), 285–305 (2006)
8. Stack, M.S., Fishman, D.A.: Ovarian Cancer, 2nd edn. Cancer Treatment and Research, 409p. Springer, New York (2009)
9. Bamberger, E., Perrett, C.: Angiogenesis in epithelian ovarian cancer (review). Molecular Pathology 55, 348–359 (2002)
10. Sprindzuk, M., Dmitruk, A., Kovalev, V., Bogush, A., Tuzikov, A., Liakhovski, V., Fridman, M.: Computer-aided image processing of angiogenic histological samples in ovarian cancer. Journal of Clinical Medicine Research 1(5), 249–261 (2009)
11. Folkman, J.: What is the evidence that tumors are angiogenesis dependent? Journal of the National Cancer Institute 82(1), 4–6 (1990)
12. Hsu, W., Lee, M., Zhang, J.: Image mining: Trends and developments. Journal of Intelligent Information Systems 19(1), 7–23 (2002)
13. Herold, J., Loyek, C., Nattkemper, T.W.: Multivariate image mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1(1), 2–13 (2011)
14. Perner, P.: Image mining: Issues, framework, a generic tool and its application to medical image diagnosis. Engineering Applications of Artificial Intelligence 15(2), 205–216 (2002)
15. Chen, W., Meerc, P., Georgescud, B., He, W., Goodellb, L.A., Forana, D.J.: Image mining for investigative pathology using optimized feature extraction and data fusion. Computer Methods and Programs in Biomedicine 79, 59–72 (2005)
16. Kovalev, V., Prus, A., Vankevich, P.: Mining lung shape from x-ray images. In: Perner, P. (ed.) MLDM 2009. LNCS, vol. 5632, pp. 554–568. Springer, Heidelberg (2009)
17. Kovalev, V., Safonau, I., Prus, A.: Histological image mining for exploring textural differences in cancerous tissue. In: Swedish Symposium on Image Analysis (SSBA 2010), March 11-12, pp. 113–116. Uppsala University, Uppsala (2010)
18. Huang, J., Kumar, R., Mitra, M., Zhu, W.J., Zabih, R.: Image indexing using color correlograms. In: IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, pp. 762–768. IEEE Comp. Soc. Press, Los Alamitos (1997)

19. Kovalev, V., Volmer, S.: Color co-occurrence descriptors for querying-by-example. In: Int. Conf. on Multimedia Modelling, Lausanne, Switzerland, Lausanne, Switzerland, pp. 32–38. IEEE Comp. Soc. Press, Los Alamitos (1998)
20. Julesz, B.: Foundations of Cyclopean Perception, p. 426. The MIT Press, Cambridge (2006)
21. Beason-Held, L.L., Purpura, K.P., Krasuski, J.S., et al.: Cortical regions involved in visual texture perception: a fMRI study. Cognitive Brain Research 7, 111–118 (1998)
22. Petrou, M., Kovalev, V., Reichenbach, J.: Three-dimensional nonlinear invisible boundary detection. IEEE Trans. Image Processing 15(10), 3020–3032 (2006)
23. Kovalev, V., Petrou, M., Suckling, J.: Detection of structural differences between the brains of schizophrenic patients and controls. Psychiatry Research: Neuroimaging 124, 177–189 (2003)

A Diffeomorphic Matching Based Characterization of the Pelvic Organ Dynamics

Mehdi Rahim[1], Marc-Emmanuel Bellemare[1], Nicolas Pirró[2], and Rémy Bulot[1]

[1] LSIS UMR CNRS 6168, Aix-Marseille University, France
mehdi.rahim@lsis.org
[2] Digestive Surgery department, Hôpital La Timone, Marseille, France

Abstract. The analysis of the behavior of the pelvic organs on dynamic MRI sequences could help to a better understanding of pelvic floor pathophysiology. The main pelvic organs (bladder, uterus-vagina, rectum) are soft-tissue organs, they undergo deformations and displacements under an abdominal strain. Moreover, the inter-patient morphological variabilities of these organs are very important. In this paper, we present a methodology for the analysis of the pelvic organ dynamics based on a diffeormorphic matching method called large deformation diffeomorphic metric mapping. It allows to define a unique contour parametrization of the pelvic organs, and to estimate the organ deformations after matching the organ shape against its initial state ($t = 0$). Some promising results are presented, where the pathology detection capability of the deformation features is analyzed through an inter-patient analysis. Also, an organ parcellation is proposed by performing a local deformation analysis.

Keywords: dynamic MRI, pelvic dynamic, shape matching.

1 Introduction

The pathologies associated with the pelvic floor disorders are characterized by an abnormally large organ descent during a strain. Besides the clinical examination, the pelvic dynamic MRI is a recommended tool for the clinical diagnosis of these pathologies [1], [2]. Thanks to its appreciable contrast, the dynamic MRI allows to qualitatively assess the behavior of the main pelvic organs (bladder, uterus-vagina, rectum) during an abdominal strain. Although significant research has been performed, the pelviperineal physiology and the anatomic basis of pelvic floor diseases remain unclear, as mentioned in [3]. An image-analysis based study could bring a quantitative characterization of the pelvic organ dynamics. As it would be automatic, it could be applied to a wide number of cases. In [4], a global characterization of the deformations of the pelvic organs has been proposed. It uses shape descriptors. This approach measures shape variations of the pelvic organs, but it does not track the local dynamics of organ specific landmarks. We propose an organ dynamics characterization based on a diffeomorphic matching. This approach allows to analyze the organ behavior locally, and to compare different organ behaviors from different MRI sequences. We use

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 469–476, 2011.

the large deformation diffeomorphic metric mapping (LDDMM) method to perform the organ shape matching. Its main advantage is that it takes into account the large organ deformations that are observed along a dynamic MRI. Thanks to the LDDMM, an intra-sequence matching is proposed to track the evolution of the contour points of an organ. We propose also an inter-sequence matching which yields a unique spatial parametrization for all the MRI sequences. The input data of the proposed approach are the contours of the organs. The segmentation process is out of the scope of this study, the MRI DICOM images were segmented by clinicians providing a set of closed contours for each of the three organs of interest, as depicted in figure 1-b. We detail in section 2 the proposed approach for the organ dynamics characterization. Some results are presented in section 3, where the pathology detection capability of the computed features is analyzed through an inter-patient analysis, in addition to a local organ analysis. Section 4 concludes the paper.

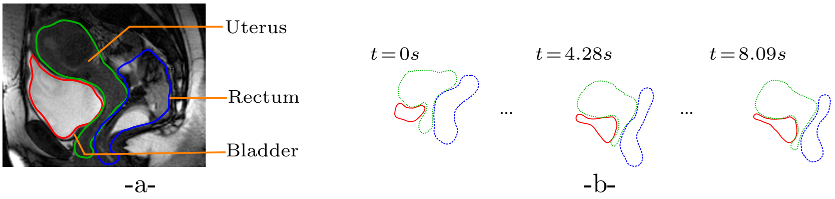


Fig. 1. a: The main pelvic organs. b: A Segmented MRI Sequence

2 Organ Dynamics Characterization

A mere visual observation of dynamic MRI sequences reveals that there is a large shape variability of the pelvic organs as they are soft-tissue organs. The shape and the size of the organs are not consistent indicators for the pelvic characterization. This is the main reason for our focus on the dynamics of the organ and its deformation evolution during a strain, compared to the shape of this organ at the rest state ($t = 0$).

Thus, for each organ type, we define a spatial reference which is common to all the studied sequences, so that we can track the points of the contours of the pelvic organs.

Our methodology is based on two matching stages :

- *Intra-sequence matching* between the current organ shape and the organ shape at the rest state, produces the deformation field of the organ points along a sequence of frames.
- *Inter-sequence matching* regarding a *template* shape calculated from the available data, provides a common spatial reference, in order to follow the same point on different sequences.

We detail in the next subsections the LDDMM method, and the followed steps for intra-sequence and inter sequence matching.

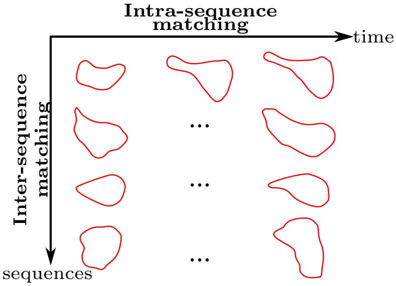


Fig. 2. The proposed matching scheme

2.1 Large Deformation Diffeomorphic Metric Mapping

About. The (LDDMM) is a non-linear registration method. It defines a matching function ϕ, which is a diffeomorphism since it is invertible and differentiable. One of the benefits of the method relies in the ability to map very irregular deformations, resulting on a smooth displacement field, regular and without intersections. The LDDMM can be applied to different geometrical structures, such as surfaces, curves, sparse points, or landmarks. More details are discussed in [5]. The method has been used, among others, for the inter-subject registration of cortical surfaces [6], the analysis of the temporal evolution of the skull [7], or an anatomical atlas estimation [8].

Mathematical Formulation. In our case, the LDDMM is applied on curves. For two curves S,T, the source and the target curves respectively, the diffeomorphic matching transforms T via a diffeomorphic function ϕ, to find a mapping between the points of the two curves.

$$\phi_1^v.T = S$$

Formula (1) defines the displacement flow v_t depending on time t.

$$\frac{\delta\phi_t^v}{\delta t} = v_t(\phi_t^v) \tag{1}$$

With : $\phi_0^v(T) = T$ and $\phi_1^v(T) = S$.

The goal of the matching is to find an optimal diffeomorphism ϕ by minimizing the formula (2) below:

$$J_{T,S}((v_t)_{t\in[0,1]}) = \gamma \int \|v_t\|_V^2 dt + \frac{1}{\sigma} E(\phi_1^v.T, S) \tag{2}$$

The first term is a regularization term that controls the smoothness of the diffeomorphism, the second term is a fidelity to data term which quantifies the matching error between the mapped target points $\mu_{\phi_1^v(T)}$ and the source points μ_S. It is defined as follows:

$$E(\phi_1^v.T, S) = \|\mu_{\phi_1^v(T)} - \mu_S\|^2 \tag{3}$$

Typically, the resolution of the problem is done with the gradient descent method. We used the implementation presented in [5].

2.2 Inter-Sequence Matching

For a given organ, the inter-sequence matching aims at determining a common contour parametrization. It involves the curve matching between an average organ shape and the organ shapes at the rest state from different sequences. For this purpose, a pre-processing step is required to obtain an ordered contour curve C where each point $c(t) = (x(t), y(t))$. The contours do not necessarily share the same number of points and the same starting point.

The inter-sequence matching is performed according to the following steps:

1. Template computation : by calculating an average initial shape, all the available organ shapes at $t = 0$ are aligned. This alignment is done according to the main geometric transformations:
   - Translation : the shapes are translated to share the same center of mass.
   - Rotation : the shapes are rotated so that the principal axes of the shapes are aligned.
   - Scaling : the normalization is performed towards a unit disk.

   The cumulative addition of the aligned shapes generates a level map (figure 3-a), from which an average shape is extracted by thresholding, typically the threshold is defined as $(max + min)/2$.
2. Shape matching : A matching by LDDMM is applied on the initial organ shapes of all the patients compared with the template shape.

Figure 3 shows consistent results.

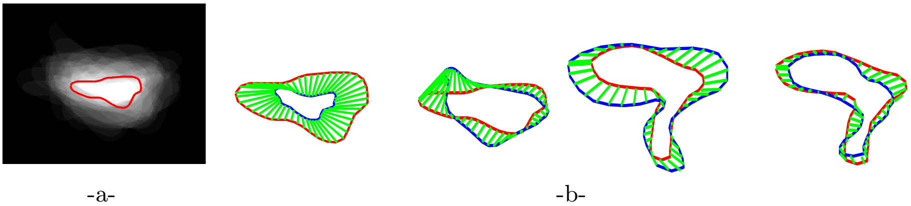


Fig. 3. a: The average shape of the bladder. b: Some results of the inter-sequence matching between different initial shapes (blue) and the associated templates (red).

2.3 Intra-sequence Matching

Thanks to the inter-sequence matching, all the contours of a given organ type (bladder, uterus or rectum) at rest share a common parametrization. So each organ contour point is clearly identified. The intra-sequence analysis aims at tracking each contour point. This is done by matching the organ points at the i^{th} frame towards the corresponding points of the initial organ shape. This matching is obtained from a composition of consecutive matching $(j, j+1)$, starting from the initial contour C_0 to the contour C_i (figure 4). This technique involves estimating a large deformation with a composition of consecutive deformations and consequently smaller ones. It avoids the inaccuracies due to very large deformations specially on the last frames of the sequence.

$$\begin{tikzcd}
C_0 \arrow[r, "\phi_{0,1}"] \arrow[rr, bend right, "\phi_{0,2} = \phi_{1,2} \circ \phi_{0,1}"'] & C_1 \arrow[r, "\phi_{1,2}"] & C_2 \cdots & C_n
\end{tikzcd}$$

Fig. 4. Diagram of the LDDMM matching on a contour sequence

3 Results and Discussion

The LDDMM provides the correspondence between the organ points. For a given sequence of m frames, the n matched points are represented with :

$$P = \begin{pmatrix} (x_1, y_1)^1 & . & (x_1, y_1)^m \\ . & (x_i, y_i)^k & . \\ (x_n, y_n)^1 & . & (x_n, y_n)^m \end{pmatrix}$$

The displacement magnitude of a point (x_i, y_i) at the k^{th} frame compared to its initial state $(t = 0)$ is defined as:

$$Def(i, k) = \sqrt{(x_i^k - x_i^1)^2 + (y_i^k - y_i^1)^2} \tag{4}$$

This measurement is the main feature of a statistical analysis of the sequences that we will develop in the next section. The analysis will involve the bladder and uterus.

3.1 Inter-subject Analysis

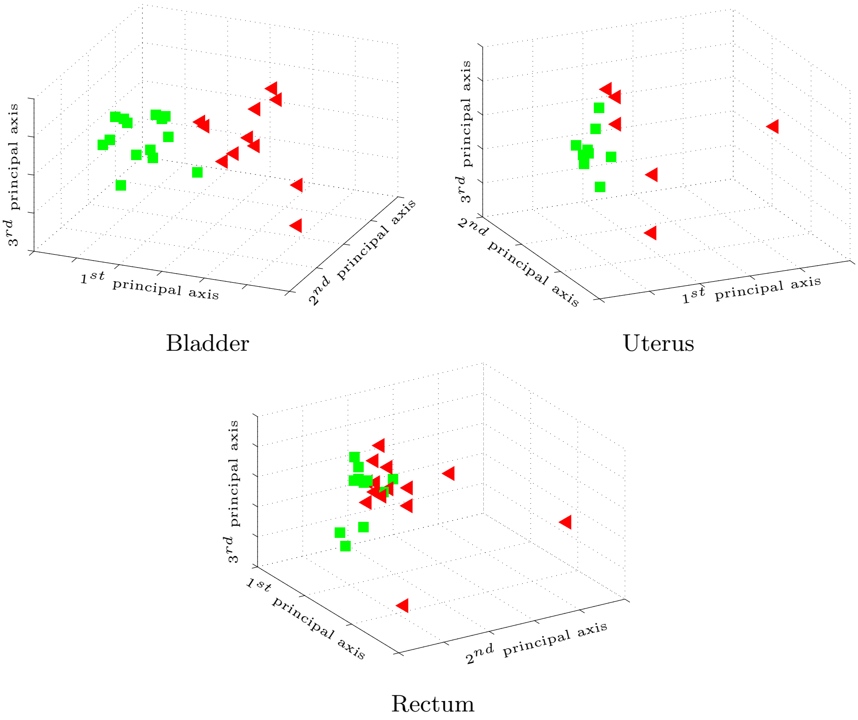
The purpose of the inter-patient analysis is to assess the ability of the deformation estimators towards the discrimination between MRI sequences. We used in this analysis a dataset consisting of 30 segmented sequences, where each sequence contains $m = 12$ frames (1 frame per second). In addition, we have the clinical diagnosis relating to each organ of the segmented pelvic MRI. The dynamics of a pelvic organ j is quantitatively summarized by a $n \times 11$ matrix of deformation magnitudes, for n contour points.

$$D_j(i, k) = \{Def(i, k + 1), 1 \leq i \leq n, 1 \leq k \leq 11\}$$

Thanks to the inter-sequence matching, we have a common and ordered set of contour points for each organ. Therefore, it is possible to compare the dynamics of two organs from two different sequences, on the basis of these points. Thus, we can rewrite the deformation matrix of a patient as a vector. All the $l = 30$ sequences data can be written as:

$$P = \begin{array}{cc} \xrightarrow{features} & \\ \begin{pmatrix} D_1(1,1) & . & D_1(n,11) \\ . & . & . \\ . & D_j(i,k) & . \\ . & . & . \\ D_l(1,1) & . & D_l(n,11) \end{pmatrix} & \downarrow patient \end{array}$$

We carried out a principal component analysis (PCA) on this matrix, in order to reduce the feature dimension. Figure 5 shows the result of the PCA, where we have represented the different sequences according to the 1^{st},2^{nd} and 3^{rd} principal component. The inertia ratio are 75% for the bladder, 81% for the uterus, 77% for the rectum, which seems sufficient to represent the sequences consistently. We observe an evident separation between the pathological cases (red triangles), and the healthy ones (green squares). This allows us to assert that the deformation magnitude calculated is a relevant criterion aiming at distinguishing in between pelvic organs those with disease and the healthy ones.



Bladder

Uterus

Rectum

Fig. 5. Inter-subject representation (green square: healthy case, red triangle : pathological case), according to the 3 PCA components

3.2 Organ Parcellation

We characterize the local deformations of the pelvic organs in order to find a parcellation related to the dynamics of the organ deformations. The purpose of the local analysis is to group the contour points according to their deformation profiles. It seems obvious to use a method of unsupervised clustering. Among the many existing clustering methods, we opted for the *Affinity propagation* method proposed in [9]. Its main advantage lies in the automatic determination of the number of clusters when this latter is not known in advance. The *Affinity Propagation* algorithm requires as input a similarity matrix built from the dataset samples. In our case, the samples are the contour points of an organ, represented by their deformations. Thus, the similarity matrix M of a contour is defined beneath:

$$M = \{m_{i,j} = d(V(i), V(j)), 1 \leq i, j \leq n\}$$
$$V(i) = \big(Def(i,1)\ ..\ Def(i,k)\ ..\ Def(i,11)\big)$$

d is the Euclidean distance. $V(i)$ is the deformation vector of the point i.

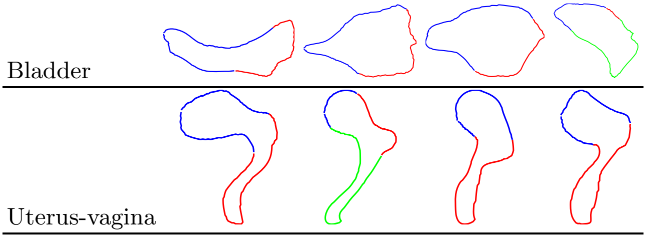


Fig. 6. Some results of the organ parcellation applied to the bladder and the uterus

From the bladder parcellation results, two to three separate sectors (blue, red, green) are distinguished. On figure 6, the blue sector corresponds to the upper edge of the bladder which undergoes less deformation than the lower edge of the bladder delimited by the red and green sectors. This result is anatomically interesting as the delimited sectors involve two anatomical references. Indeed, the point undergoing the maximal deformation corresponds to an anatomical landmark called the bladder neck, the blue parcel includes the attachment point of the bladder to the urachus, while the red parcel includes the bladder neck.

Two to three sectors characterize the uterus, they separate the uterus from the vagina. In addition, the deformation magnitude of the sectors related to the vagina are larger than their uterus counterparts, corroborating the fact that the uterus is more rigid than the vagina. Globally, the local analysis helped to highlight automatically the non-homogeneous deformations of the pelvic organ, and to delimit these non-homogeneous sectors which have a clinical meaning.

4 Conclusion

The LDDMM allowed us to solve the problem of matching the contours of organs which have mobile anatomical landmarks. The intra-sequence and inter-sequence matching defined a geometrical reference of the pelvic organs, despite of the morphological variability of the soft-tissue organs.

We have analyzed several pelvic dynamic MRI sequences, by estimating the magnitude of the deformations undergone by these pelvic organs during a strain. The global statistical analysis of shape deformations helped to distinguish the pathological cases from the healthy ones. This result validates the relevance of the features chosen for the pelvic dynamics assessment. A study on a larger dataset of MRI sequences would open the possibility to build a diagnosis-aid system.

The local analysis of the bladder and the uterus highlighted the anisotropic deformation properties of these organs, which cannot be retrieved by common geometrical features, such as the area, the perimeter, etc..

As a result of the local analysis, we proposed a parcellation of the organ contour. It is based on the deformation profiles of the organ points. It provides a delimitation of parcels associated with significant anatomical references.

On the whole, this analysis contributes to a better understanding of the dynamics of the pelvic organs. Moreover, it will be a key feature for the validation of the biomechanical behavior laws of the pelvic organs, used in simulations [10]. Indeed, we are able to compare local organ deformations which are observed with MRI to simulated ones. Furthermore, in order to have a fully automated process, we are currently working on automating the segmentation, using a contour tracking method.

Acknowledgments. We thank J. Lefèvre for the fruitful discussions about the LDDMM method. This work is supported by the French National Research Agency (ANR) under reference "ANR-09-SYSC-008".

References

1. Fielding, J.R.: Mr imaging of pelvic floor relaxation. Radiologic Clinics of North America 41(4), 747–756 (2003)
2. Seynaeve, R., Billiet, I., Vossaert, P., Verleyen, P., Steegmans, A.: MR imaging of the pelvic floor. JBR-BTR 89(4), 182–189 (2006)
3. Weber, A.M., Richter, H.E.: Pelvic organ prolapse. Obstetrics and Gynecology 106(3), 615–634 (2005)
4. Rahim, M., Bellemare, M.E., Pirró, N., Bulot, R.: A shape descriptors comparison for organs deformation sequence characterization in mri sequences. In: IEEE International Conference on Image Processing, ICIP 2009, pp. 1069–1072 (2009)
5. Glaunes, J., Qiu, A., Miller, M., Younes, L.: Large deformation diffeomorphic metric curve mapping. Int. Journal of Computer Vision 80(3), 317–336 (2008)
6. Auzias, G., Glaunès, J., Cachia, A., Cathier, P., Bardinet, E., Colliot, O., Mangin, J., Trouve, A., Baillet, S.: Multi-scale diffeomorphic cortical registration under manifold sulcal constraints. In: IEEE International Symposium on Biomedical Imaging–ISBI 2008, pp. 1127–1130 (2008)
7. Durrleman, S., Pennec, X., Trouvé, A., Gerig, G., Ayache, N.: Spatiotemporal atlas estimation for developmental delay detection in longitudinal datasets. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009. LNCS, vol. 5761, pp. 297–304. Springer, Heidelberg (2009)
8. Beg, M., Khan, A.: Computing an average anatomical atlas using LDDMM and geodesic shooting. In: IEEE International Symposium on Biomedical Imaging–ISBI 2006, pp. 1116–1119 (2006)
9. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science 315, 972–976 (2007)
10. Bellemare, M.E., Pirró, N., Marsac, L., Durieux, O.: Toward the simulation of the strain of female pelvic organs. In: IEEE EMBS Annual International Conference, pp. 2756–2759 (2007)

Histogram-Based Optical Flow for Functional Imaging in Echocardiography

Sönke Schmid[1,2,⋆], Daniel Tenbrinck[1,2,⋆], Xiaoyi Jiang[1,2],
Klaus Schäfers[2], Klaus Tiemann[3], and Jörg Stypmann[2,3]

[1] Dept. of Mathematics and Computer Science, University of Münster, Germany
[2] European Institute for Molecular Imaging, Münster, Germany
[3] Dept. of Cardiology and Angiology, University Hospital of Münster, Germany

Abstract. Echocardiographic imaging provides various challenges for medical image analysis due to the impact of physical effects in the process of data acquisition. The most significant difference to other medical data is its high level of speckle noise that makes the use of conventional algorithms difficult. Motion analysis on ultrasound (US) data is often referred to as 'Speckle Tracking' which plays an important role in diagnosis and monitoring of cardiovascular diseases and the identification of abnormal cardiac motion. In this paper we address the problem of speckle noise within US images for estimating optical flow. We demonstrate that methods which directly use image intensities are inferior to methods using local features within the US images. Based on this observation we propose an optical flow method which uses histograms as a local feature of US images and show that this approach is more robust under the presence of speckle noise than classical optical flow methods.

Keywords: Ultrasound, Motion Analysis, Optical Flow, Histogram.

1 Introduction

Motion analysis in echocardiography is a fundamental part in diagnosis of cardiovascular diseases. By tracing the endocardial border of the myocardium physicians assess different medical parameters. Based on these measurements abnormal motion of the myocardium can be identified and quantified, hence helping in computer aided diagnosis [9]. Recently, optical flow methods have been proposed for estimating motion in echocardiographic data [1,3,4]. These methods deliver a dense motion field in contrast to contour-based algorithms currently used in clinical environment.

In Section 2 we analyze the noise model of ultrasound images, which is of multiplicative nature. Based on this observation we will show that the fundamental assumption of most optical flow methods, the so called 'Intensity Constancy Constraint' (ICC), is heavily violated in the presence of multiplicative noise in US imaging. To deal with this problem we propose local cumulative histograms

[⋆] The authors contributed equally to this work.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 477–485, 2011.

as a discrete representation of the intensity distribution in the neighbourhood of a pixel. We incorporate this feature into a new basic constraint in Section 3 and propose a novel optical flow algorithm based on this constraint. To show the robustness of our approach in the presence of multiplicative noise we present in Section 4 results on synthetic (ultrasound software phantom) as well as real patient data from echocardiographic examinations and compare our method with a popular respresentative of optical flow methods based on the ICC. Finally, this paper is concluded by discussion in Section 5.

2 Motivation

2.1 Optical Flow and Constancy Constraints

Optical flow methods are a popular approach to compute motion between two given images. These methods model motion as a dense vector field and allow to incorporate a-priori knowledge into estimation of the motion. Recently, different optical flow methods have been proposed for medical imaging [1,3]. Generally, these algorithms are based on the fundamental assumption that for a motion vector (u, v) the intensity of two pixels in images at time t and $t+1$ is constant:

$$I(x, y, t) = I(x + u, y + v, t + 1) \ . \tag{1}$$

Equation 1 is referred to as the Intensity Constancy Constraint.

The ICC implies that the illumination does not change and no noise is present in the images. On real data the influence of noise can be alleviated by smoothing the images. However, in ultrasound imaging this procedure does not provide good results due to the high level of speckle noise which will be discussed in the next section.

In the literature there are also other constancy constraints. Examples are gradient, Hessian, and Laplacian of corresponding pixels [2,7]. All such constraints share the same problem as ICC when dealing with ultrasound data.

2.2 Speckle Noise

One has to carefully deal with speckle noise in ultrasound imaging. The origins of speckle are tiny inhomogenities within the tissue which reflect ultrasound waves but cannot be resolved by the ultrasound system. Speckle noise is a known phenomenon and depends on the underlying signal intensity. The image degradation process [6] can be modeled by:

$$\tilde{I}(\boldsymbol{x}) = I(\boldsymbol{x}) + s_\sigma(\boldsymbol{x}) \cdot \sqrt{I(\boldsymbol{x})} \tag{2}$$

Here, I is the unbiased image, s is Gaussian distributed random noise with mean 0 and standard deviation σ, and $\tilde{I}$ is the observed image. This multiplicative noise leads to distortions in the image, especially in regions with high intensities. In order to deal with multiplicative noise one has to use more sophisticated methods

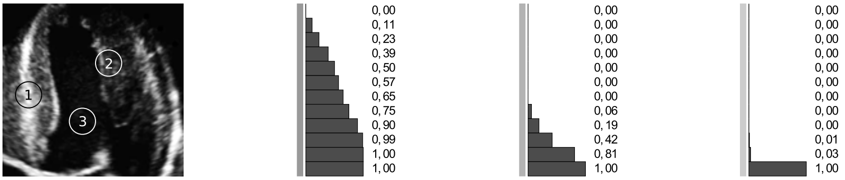
Table 1. Mean distance of two pixel patches with uniform intensity contaminated by speckle noise

Noise level (σ)	0.01	0.02	0.04	0.08	0.16
Intensity L_1 Distance	14.7335	20.8151	29.4815	41.7718	59.1326
Intensity L_2 Distance	180.1315	253.9759	360.3631	510.5504	722.3002
Histogram L_1 Distance	1.3149	1.3017	1.3050	1.3020	1.3313
Histogram L_2 Distance	0.2917	0.2895	0.2899	0.2895	0.2952

[6] compared to additive noise which is not signal dependent. The intensities of speckles can change due to motion of the imaged object, especially if the object moves out of the imaging plane in 2D ultrasound B-mode imaging. To illustrate the effect of multiplicative noise on the ICC, Table 1 shows experimental results using a 10 × 10 pixel patch with a greyscale value of $I = 120$. We added speckle noise based on Equation 2. The noise variance σ was increased to investigate the violation of the ICC by multiplicative noise. To show the impact of speckle noise we considered different distance measures L_1 and L_2. For each noise variance σ we generated $10,000$ pairs of patches and computed the mean distances. As can be seen in Table 1, the average distance on the 10 × 10 pixel patches grows significantly with increasing noise variance, indicating that the ICC and its related variants are not suitable for ultrasound imaging and we need to find more robust image features.

2.3 Local Cumulative Histograms

In contrast to the constancy assumptions conventionally used for optical flow estimation [2,7] which are all based on the pixel intensity or its related variants, we suggest a constancy constraint incorporating information from a small neighbourhood around the pixel. We propose the use of local histograms as a discrete representation of the intensity distribution to relate corresponding pixels of ultrasound images. By this way we capture all important information of this region in local histograms. We use cumulative histograms which are more robust than normal histograms for comparison purposes. The histograms are normalized so that the highest value becomes one.



(a) US B-mode image (b) Histogram 1 (c) Histogram 2 (d) Histogram 3

Fig. 1. (a) Different regions within an US image of the left ventricle. (b) Histogram of septum. (c) Histogram in region with shadowing effects. (d) Histogram of blood pool.

Figure 1 shows different local cumulative histograms of a real 2D US B-mode image using twelve bins to represent the greyscale distribution. As one can see, the three cumulative histograms can be clearly distinguished. This gives us a reliable feature especially in regions with low contrast, as can be seen in histogram 2. The three example histograms represent different regions of the cardiac image: the high intensity values of the septum (1), a mixed signal distribution in the lateral wall of the myocardium due to shadowing effects (2), and the non-reflecting blood within the myocardium (3).

3 Histogram-Based Optical Flow Method

As shown in the last section the intensity constancy constraint is not a good choice in the presence of speckle noise. Due to this fact we replace the ICC by another constraint based on the assumption that the local intensity distribution, i.e. the local cumulative histogram, remains constant over time. To exploit the effect of this replacement in detail we apply this exchange to the basic optical flow algorithm of Horn-Schunck (HS) since its properties are well-understood [5]. Note that the goal of our current work is to explore the fundamental potential of histogram-based features for motion estimation in US imaging. For this purpose it is feasible to adapt the baseline algorithm of HS. We will extend to more sophisticated optical flow estimation paradigms in future.

3.1 Histogram Constancy Constraint

We replace Equation 1 by a histogram constancy constraint (HCC):

$$H(x, y, t) = H(x + u, y + v, t + 1) \tag{3}$$

where the function H represents the local histogram of the region surrounding the pixel (x, y) at the given point of time. As shown in Figure 1 histograms can be represented by a vector whose dimension corresponds to the number of bins in the histogram. To measure the distance of the histogram vectors we use the L_2-norm. Compared to the classical ICC this constraint remains robust under the influence of multiplicitative noise, as can be seen in Table 1.

3.2 Energy Functional

The HCC formulated as an energy functional leads to a minimization problem:

$$\operatorname*{argmin}_{u,v} \iint \|H(x + u, y + v, t + 1) - H(x, y, t)\|^2 \, dxdy \ . \tag{4}$$

Without additional regularization this problem is ill-posed. The smoothness of the flow field (u, v) is a reasonable regularization for the presented application due to the fact that human tissue can be deformed up to a certain level but cannot change its topology. The HCC formulated as an energy functional combined with the smoothness constraint leads to the final minimization problem:

$$\underset{u,v}{\operatorname{argmin}} \iint \|H(x+u, y+v, t+1) - H(x, y, t)\|^2 + \alpha \left(|\nabla u|^2 + |\nabla v|^2 \right) dxdy \quad (5)$$

where α is the smoothness parameter regulating the influence of the smoothness constraint.

3.3 Numerical Discretization

The minimization problem given by the Equation 5 is solved numerically with use of the Euler-Lagrange Theorem. We replace the HCC by its Taylor-approximation to the first order:

$$H(x+u, y+v, t+1) - H(x, y, t) \approx H_x(x, y, t) \cdot u + H_y(x, y, t) \cdot v + H_t(x, y, z) \quad (6)$$

The derivatives H_x, H_y, and H_t of the histograms can be approximated by the finite differences:

$$\begin{aligned} H_x(x, y, t) &= (H(x+1, y, t) - H(x-1, y, t)) \,/\, 2 \\ H_y(x, y, t) &= (H(x, y+1, t) - H(x, y-1, t)) \,/\, 2 \\ H_t(x, y, t) &= (H(x, y, t+1) - H(x, y, t)) \end{aligned} \quad (7)$$

With the Taylor-approximated HCC the Euler-Lagrange equations of the minimization problem are given by:

$$\begin{aligned} (H_x * H_x)\, u + (H_x * H_y) v + (H_x * H_t) &= \alpha \Delta u = \alpha\, (\overline{u} - u) \\ (H_y * H_x)\, u + (H_y * H_y) v + (H_y * H_t) &= \alpha \Delta v = \alpha\, (\overline{u} - u) \end{aligned} \quad (8)$$

where the operator '$*$' represents the scalar product of two vectors. This linear system can be solved directly for u and v and leads to an iterative scheme for computing the histogram-based optical flow (HOF):

$$\begin{aligned} u &= \frac{(\alpha + H_y * H_y) \cdot (\alpha \overline{u} - H_x * H_t) - (H_x * H_y) \cdot (\alpha \overline{v} - H_y * H_t)}{(\alpha + H_x * H_x)\,(\alpha + H_y * H_y) - (H_x * H_y)^2} \\ v &= \frac{(-H_x * H_y) \cdot (\alpha \overline{u} - H_x * H_t) + (\alpha + H_x * H_x) \cdot (\alpha \overline{v} - H_y * H_t)}{(\alpha + H_x * H_x)\,(\alpha + H_y * H_y) - (H_x * H_y)^2} \end{aligned} \quad (9)$$

3.4 Implementation

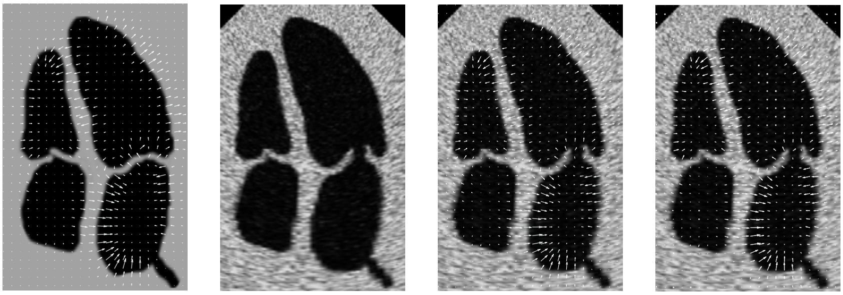
The histogram-based optical flow (HOF) was implemented for 2D images and also 3D data (with the necessary extensions) from newest ultrasound systems. To cope with large movements we extended our method by a simple multiscale approach as described in [7].

The selection of function H has to be considered carefully depending on the type of data, since there is a tradeoff between more statistics in larger regions and loss of locality. For ultrasound images we found a mask size of 9×9 pixels as best choice in combination with a Gaussian weighting function to give the pixels in the center a higher influence on the histogram. For the discretization of the intensity distribution we used histograms with 30 bins which proved to be fully sufficient. Since the L_2 distance of two normalized histogram vectors is much smaller than the difference of the intensity values used in HS the smoothness parameter α has to be chosen accordingly smaller. For ultrasound data empirical tests on 15 datasets showed optimal values for α in the domain $\alpha \in [0.5, 1.5]$ in contrast to $\alpha \in [200, 5000]$ for HS. This specification is bounded to the design parameters stated above (number of bins, mask size).

4 Results

4.1 Software Phantom

In order to validate our method quantitatively we used the simulation proposed in [8] to create a software phantom for US. Simulating the data acquisition by an US transducer this method allows to add speckle noise to a given image and also includes deformation effects. Thus, two given images with a known ground truth motion can be transformed into a software phantom to validate optical flow algorithms on US data (see Figure 2). The simulation software contains several parameters that were selected as suggested by medical experts. For evaluation we optimized the parameters of both HS and HOF algorithm and compared the results by using the average endpoint error [2]. As can be seen in Table 2, our new constraint improves the motion estimation significantly. Tests on five additional datasets showed very similar results. Although the absolute difference of performance does not seem to be large, an improvement of 28% has been



(a) ground truth (b) speckle phantom (c) HOF result (d) HS result

Fig. 2. Synthetic data simulating a four-chamber view of the heart. (a) Original image with ground truth flow. (b) Speckle phantom. (c) Result of the HOF algorithm. (d) Result of the HS algorithm.

Table 2. Comparison of the performance of the HOF-algorithm to the method of Horn-Schunck using the average endpoint error [2] and its standard deviation in pixel

Sequence	HOF	HS
Software phantom with multiscale	0.387 ± 0.280	0.455 ± 0.310
Software phantom without multiscale	0.414 ± 0.290	0.530 ± 0.362

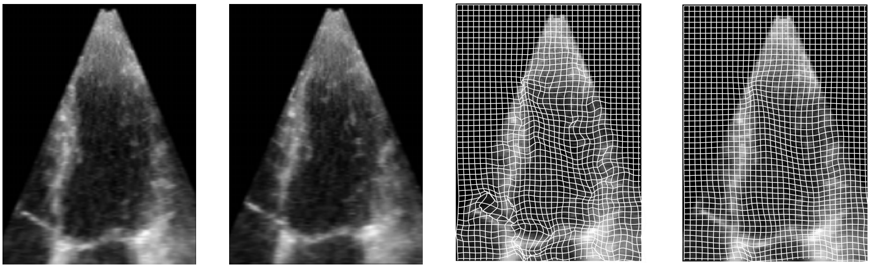

(a) floating frame (b) target frame (c) HS result (d) HOF result

Fig. 3. (a),(b) Consecutive US B-mode images of the left ventricle. (c) Result of Horn-Schunck algorithm. (d) Result of our approach based on local histograms.

reached just by incorporation of a new basic constraint into the algorithm of HS. Furthermore, the standard deviation is also reduced by our approach. We expect further improvements if the HCC gets used with more sophisticated methods from the literature.

4.2 2D Ultrasound B-Mode Images

To validate our approach on real medical data we chose several 2D US B-mode images of the left ventricle acquired with a Philips iE33 ultrasound system. In Figure 3(a) one can see a real image of the left ventricle in a four-chamber view. We chose two consecutive frames in systole of the myocardium for motion estimation. Deformation grids were used to visualize the estimated motion vectors and we let experts in echocardiography rate the quality of these estimations. Figure 3(c) shows a result of the Horn-Schunck algorithm with the regularization parameter $\alpha = 250$. The visualization by a deformed grid reveals several inconsistencies and anatomically incorrect deformations although we chose a relatively high regularization. This is due to the fact that the HS algorithm is based on the ICC, which is not valid in the presence of speckle noise as described in Section 2.2. In Figure 3(d) we demonstrate the result of our approach for $\alpha = 1$. One can clearly see that the proposed HCC leads to satisfying results on noisy US images with low regularization, though there is motion out of the image plane which cannot be estimated on 2D images.

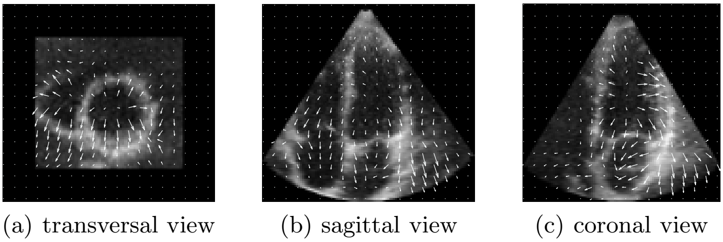

(a) transversal view (b) sagittal view (c) coronal view

Fig. 4. Transversal, sagittal, and coronal slices of an US 3D data set. The vectors indicate the result of motion estimation with HOF.

4.3 3D Echocardiographic Data

We also tested our algorithm on real 3D data from an echocardiographic TEE examination of the left ventricle captured with a Siemens iE33 ultrasound system. Figure 4 illustrates the results of motion estimation for a 3D dataset consisting of $112 \times 104 \times 104$ voxels. It shows three orthogonal slices of the dataset with the corresponding motion vectors in sagittal, coronal, and transversal slices. Since the full motion of the left ventricle can be captured in the volume dataset, less problems occurr in the estimation of the flow fields. The additional dimension enforces anatomically consistent flow fields even more. For this reason we chose the regularization parameter $\alpha = 0.6$ and observed satisfying results which give anatomically consistent flow fields in all three dimensions. Thus, our method can be used for functional imaging with 3D ultrasound data which is a new and fast developing field in clinical environment.

5 Discussion

We proposed a new constraint for optical flow methods targeted for the use on echocardiographic data. Our approach is based on local cumulative histograms which have shown to be robust in the presence of speckle noise in ultrasound imaging. It was shown by using the Horn-Schunck algorithm as a representative example that ignoring this fundamental fact tends to produce bad motion estimations and inconsistent flow fields. In future we will extend our work to incorporate the proposed HCC into more sophisticated methods from the literature. Furthermore, we want to develope a validation method for functional imaging by using a hardware phantom for motion simulation.

Acknowledgments. This study was supported by the Deutsche Forschungsgemeinschaft (DFG), SFB 656 MoBil, Münster, Germany (project C3) and the IZKF Münster Core Unit "Small animal ultrasound: imaging and therapy" (ECHO).

References

1. Achmad, B., Mustafa, M., Hussain, A.: Inter-frame Enhancement of Ultrasound Images Using Optical Flow. In: Badioze Zaman, H., Robinson, P., Petrou, M., Olivier, P., Schröder, H., Shih, T.K. (eds.) IVIC 2009. LNCS, vol. 5857, pp. 191–201. Springer, Heidelberg (2009)
2. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A Database and Evaluation Methodology for Optical Flow. Int. J. Comput. Vis. 92, 1–31 (2011)
3. Dawood, M., Büther, F., Jiang, X., Schäfers, K.: Respiratory Motion Correction in 3D PET Data with Advanced Optical Flow Algorithms. IEEE Trans. Medical Imaging 9, 1164–1175 (2008)
4. Duan, Q., Angelini, E., Lorsakul, A.: Coronary Occlusion Detection with 4D Optical Flow Based Strain Estimation on 4D Ultrasound. In: Proc. of 5th Int. Conf. on Functional Imaging and Modeling of the Heart, pp. 211–219 (2009)
5. Horn, B., Schunck, B.: Determining Optical Flow. Artificial Intelligence 17, 185–203 (1981)
6. Jin, Z., Yang, X.: A Variational Model to Remove the Multiplicative Noise in Ultrasound Images. J. Math. Imaging Vis. 39, 62–74 (2011)
7. Papenberg, N., Bruhn, A., Brox, T.: Highly Accurate Optic Flow Computation with Theoretically Justified Warping. Int. J. Comput. Vision 67, 141–158 (2006)
8. Perreault, C., Auclier-Fortier, M.-F.: Speckle Simulation Based on B-Mode Echographic Image Acquisition Model. In: Proc. of 4th Canad. Conf. Comp. and Robot Vis., pp. 379–386 (2007)
9. Stypmann, J., Engelen, M., Troatz, C., Rothenburger, M., Eckardt, L., Tiemann, K.: Echocardiographic assessment of global left ventricular function in mice. Laboratory Animals 43(2), 127–137 (2009)

No-reference Quality Metrics for Eye Fundus Imaging

Andrés G. Marrugo[1], María S. Millán[1], Gabriel Cristóbal[2], Salvador Gabarda[2], and Héctor C. Abril[1]

[1] Group of Applied Optics and Image Processing, Department of Optics and Optometry, Universitat Politècnica de Catalunya, Terrassa, Spain
{andres.marrugo,hector.abril}@upc.edu, millan@oo.upc.edu
[2] Instituto de Óptica "Daza de Valdés" (CSIC), Serrano 121, Madrid 28006, Spain
{gabriel,salvador}@optica.csic.es

Abstract. This paper presents a comparative study on the use of no-reference quality metrics for eye fundus imaging. We center on auto-focusing and quality assessment as key applications for the correct operation of a fundus imaging system. Four state-of-the-art no-reference metrics were selected for the study. From these, a metric based of Rényi anisotropy yielded the best performance in both auto-focusing and quality assessment.

Keywords: No-reference metrics, fundus image, image quality.

1 Introduction

Eye fundus imaging is an integral part of modern ophthalmology, and as so it can truly benefit from emerging methods for image content estimation and quality assessment. In this paper we present a preliminary study on the use of no-reference measures of image content in fundus imaging. We have chosen four state-of-the-art no-reference metrics that have been recently introduced. In the following sections we discuss the applicability of these metrics in two different aspects of fundus imaging: auto-focusing and image quality assessment. In fact, most of the no-reference quality assessment methods were initially proposed in the context of autofocusing applications [1]. These two aspects play a crucial role in the correct operation of a fundus imaging system, which at present day are still chiefly performed by human operation.

2 No-reference Metrics

No-reference assessment of image content is, perhaps, one of the most difficult –yet conceptually simple– problems in the field of image analysis [2]. It is only until recently that several authors have proposed no-reference metrics in an attempt to shed some light on this uncertain problem. We have considered four metrics to apply them in fundus imaging. The first metric Q_1 was proposed by Gabarda

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 486–493, 2011.

and Cristóbal [3] and is based on measuring the variance of the expected entropy of a given image upon a set of predefined directions. The entropy is computed on a local basis using the generalized Rényi entropy and the normalized pseudo-Wigner distribution as an approximation for the probability density function. Therefore, a pixel-by-pixel entropy can be computed, and histograms as well. The Rényi entropy associated to a pixel n in an image can be computed as:

$$R[n] = -\frac{1}{2}\log_2\left(\sum_{k=1}^{N} \breve{P}_n^3[k]\right) \quad , \tag{1}$$

where N is the size of the spatial window used, and $\breve{P}$ is the normalized probability distribution. We can now compute an entropy value for any given orientation θ_i to obtain $R[n, \theta_i]$. The expected value for the whole image is calculated as:

$$\bar{R}[\theta_i] = \sum_n R[n, \theta_i]/M \quad , \tag{2}$$

where M is the image size. And finally the standard deviation from the expected entropy for K orientations –the metric itself– is computed as:

$$Q_1 = \left(\sum_{i=1}^{K} \left(\mu - \bar{R}[\theta_i]\right)^2 /K\right)^{1/2} \quad , \tag{3}$$

where μ is the mean of $\bar{R}[\theta_s]$. Q_1 is a good indicator of anisotropy and the authors were able to show that this measure provides a good estimate for the assessment of fidelity and quality in natural images, because their degradations may be seen as a decrease in their directional properties. This directional dependency is also true for fundus images, especially due to blurring or uneven illumination

A drawback of Q_1 is that it requires uniform degradation across the whole image. However, here we show that the use of domain knowledge for retinal imaging provides a means to adjust the metric so as to meet local quality requirements. For this case it would imply to multiply every $R[n, \theta_s]$ by a weighting function $w[n] \in [0, 1]$ such that some specific areas are given more importance,

$$\bar{R}[\theta_i] = \sum_n R[n, \theta_i]w[n]/M \quad . \tag{4}$$

This yields a modified Q_1'. Considering two of the most relevant features of a fundus image, the optic disc (OD) and the blood vessels, we have designed a weighting function that takes this fact into account. It is known that in order to assess image sharpness, specialists fixate on the surroundings of the OD to visualize the small blood vessels [4]. The weighting function used is an elliptic paraboloid centered at the OD with values ranging from one exactly at the position of the OD to approximately zero very near the periphery. This function has also been used to model the illumination distribution in fundus images. The approximate position of the OD is determined via template matching [5]. The spatial distribution of the weighting function is shown in Fig. 1(b).

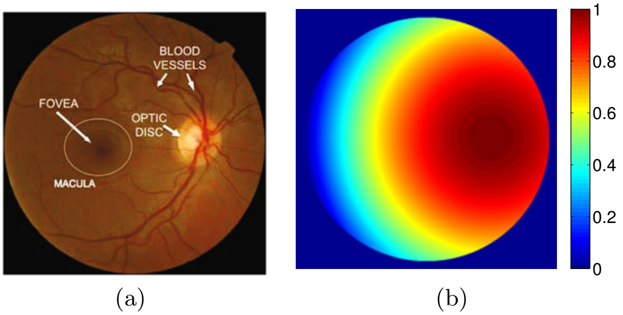


Fig. 1. (a) Normal fundus image. (b) Weighting function $w[n]$ described by an elliptic paraboloid centered at the OD.

The second metric Q_2 was recently proposed by Zhu and Milanfar [6] and it seeks to provide a quantitative measure of –what they call– "true image content". It is correlated with the noise level, sharpness, and intensity contrast manifested in visually salient geometric features such as edges. Q_2 is based upon singular value decomposition of local image gradient matrix. Its value generally drops if the variance of noise rises, and/or if the image content becomes blurry. To avoid regions without edges this algorithm divides the image into small patches and only processes anisotropic ones (non-homogeneous), thus local information is embedded into the final result.

The third metric Q_3 was proposed by Ferzli and Karam [1]. It is a sharpness metric designed to be able to predict the relative amount of blurriness in images regardless of their content. Q_3 is conceived based on the notion that the human visual system is able to mask blurriness around an edge up to a certain threshold, called the"just noticeable blur" (JNB). It is an edge-based sharpness metric based on a human visual system model that makes use of probability summation over space. JNB can be defined as the minimum amount of perceived blurriness given a contrast higher than the "Just Noticeable Difference". The probability of blur detection (P_{blur}) at an edge given a contrast C can be modeled as a psychometric function given by:

$$Q_3 = P_{blur} = P(e_i) = 1 - \exp\left(-\left|w(e_i)/w_{JNB}(e_i)\right|^{\beta}\right) \quad , \tag{5}$$

where $w_{JNB}(e_i)$ is the JNB edge width which depends on the local contrast C, $w(e_i)$ is the measured width of the edge e_i inside a small patch of the image and β is a fitting constant with a median value of 3.6. Finally, for the sake of completeness we include the image variance as metric Q_4 defined as:

$$Q_4 = \sum_n \left(I[n] - \bar{g}\right)^2 \quad , \tag{6}$$

where $I[n]$ indicates the gray level of pixel n, and $\bar{g}$ the gray mean of the image. This measure has been proven to be monotonic and has a straight-forward relation with image quality for autoregulative illumination intensity algorithms [7].

3 Experimental Details

All images were acquired using a digital fundus camera system (TRC-NW6S, Topcon, Tokyo Japan) with a Fuji FinePix S2 Pro camera, with an image resolution of 1152 × 768. The images were digitized in color RGB of 24 bit-depth in TIFF format without compression. In all figures the images are shown in color, however all metrics were computed using only the luminance channel (Y) of the YUV color space as usual in image quality assessment. From Fig. 1(a) it is evident that the region of interest of the image is that of an approximately circular shaped area that corresponds to the captured object field. The remaining black pixels are not of interest, thus all metrics have been modified to solely include pixels within the circular region of interest in the calculation. The neighboring pixels of the sharp black edge are also left aside from all calculations.

4 Fundus Auto-focusing

In fundus photography, the task of fine focusing the image is demanding and lack of focus is quite often the cause of suboptimal photographs [4]. Autofocus algorithms have arisen from the possibility that digital technology offers to continuously assess the sharpness of an image an indicate when the best focus has been achieved. Any given focus measure should be in principle monotonic with respect to blur and robust to noise.

The first experiment we carried out was to observe the behavior of the considered metrics with artificially blurred fundus images (Fig. 2(c)-(d)). Notice the detail from the sharp image and how the fine structures are properly resolved. The increase in blurriness hinders this level of detail, thus the medical use as well. In Fig. 3(a) the original sharp image (Fig. 2(a)) was convolved with a 15 × 15 Gaussian kernel with a varying standard deviation σ. All metrics are in relative value. The figure clearly reveals the overall monotonic nature of all metrics, however Q_1 is the only metric that rapidly decreases with respect to increase in blurriness.

To validate experimentally this result we captured a series of fundus images from an optimal position of focus to the end of the fine focus capability of the retinal camera (Fig. 2(e)-(f)). The fine focus knob of the retinal camera is operated manually and is able to compensate over a range of $-13 \sim +12D$. Fig. 3(b) shows the relative values for all metrics for seven images with increasing levels of blurriness. Notice how Q_1 also behaves in a consistent way with respect to the deviation from optimal focus. The other metrics seem to be reliable for a small amount of blurriness. One possible explanation for the discrepancy between the artificial and real blur for the metrics Q_{2-4} is that the overall illumination distribution cannot be exactly the same, moreover it is also non-uniform. If the metric is not conceived for variations in illumination –even if they are small– it might be prone to produce an unreliable measure. The algorithm Q_1 is based on a normalized space-frequency representation of the image and not in the image-levels statistics, hence it is robust against illumination changes. In addition, we

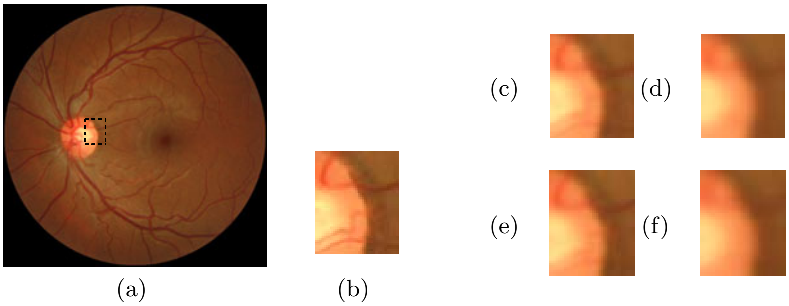


Fig. 2. (a) Original sharp fundus image and (b) detail. (c)-(d) details from artificially blurred images with σ of 1.5 and 3, respectively. (e)-(f) detail from images with different degrees of focus 3 and 6, respectively (See Fig. 3).

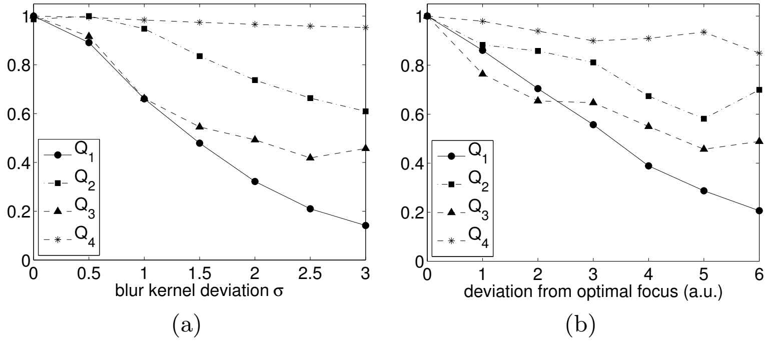


Fig. 3. No-reference metrics for assessing optimal focus in relative value. (a) Fundus image artificially blurred with a 15 × 15 gaussian kernel with varying σ. (b) Fundus images corresponding to the same eye in (a) but with different degrees of fine focus acquired with the retinal camera.

have adjusted the metric to meet the local quality requirements by means of a spatially-variant weighting function defined after the geometry of the problem. Similar results have been obtained for other fundus images (10>), but are not reported here for a matter of space.

5 Fundus Image Quality Assessment

Initial image quality is a limiting factor for automated retinopathy detection [8]. The imaging procedure is usually carried out in two separate steps: image acquisition and diagnostic interpretation. Image quality is subjectively evaluated by the person capturing the images and they can sometimes mistakenly accept a low quality image [9]. A recent study by Abràmoff et al. [10] using an automated system for detection of diabetic retinopathy found that from 10 000 exams 23%

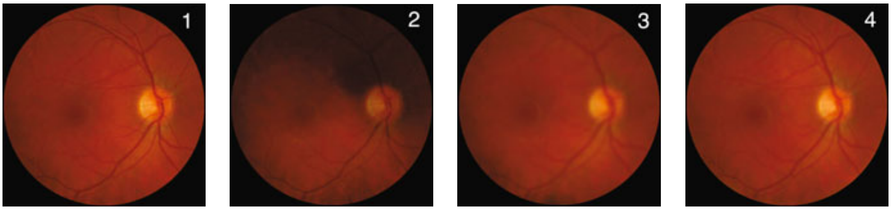


Fig. 4. Fundus images with varying degree of quality corresponding to the same eye

had insufficient image quality. Accurate quality assessment algorithms can allow operators to avoid poor images. Furthermore, a quality metric would permit the automatic submission of only the best images if many are available. It is from this point of view on that no-reference metrics can be truly useful.

It is often the case that for a given patient several fundus images are acquired. A multilevel quality estimation algorithm at the first few levels has to determine if the images correspond to fundus images, if they are properly illuminated, etc; in other words, if they meet some minimum quality and content requirements. This is in some way what the operator does, he acquires the image and then decides to accept it or not by rapidly visualizing a downscaled version of the image. Once several images of acceptable quality pass this first filter (human or machine), the system would need a final no-reference metric to decide which image to save or to send for further diagnostic interpretation. This metric should in principle yield the sharpest image, with less noise and with the most uniform illumination as possible.

Here we seek to elucidate the possible use of the no-reference metrics for fundus image quality assessment. For this experiment we have analyzed a set of 20 fundus images divided in 5 subsets of 4 images corresponding to the same eye and acquired within the same session. All images within each subset have a varying degree of quality similar to the first subset shown in Fig. 4. Our purpose is to attempt to organize this set from the best image down to the worse. The relative values from all the metrics applied to this set are shown in Table 1. Notice the value Q_1' for image 2. This image is in focus, however it suffers from uneven illumination. Q_1' puts more emphasis on the retinal structures, which are well defined in spite of the illumination, hence the increase with respect to Q_1. Illumination problems are less difficult to compensate as opposed to blurring [11]. This is in line with the specialist's evaluation of the images.

To validate the results two optometrists were recruited as readers. They were familiarized with fundus images and were asked to examine the whole set of images (4 per subject). They evaluated each subset and organized the images from the best to the worse in terms of sharpness and visibility of retinal structures. The relative scores of the metrics are converted to sorting or permutation indexes so as to compare with the quality sorting carried out by the readers (Table 2). Note that in this case only Q_1 and Q_1' agree entirely with the readers. To quantify the agreement we devised a similarity score based on the Spearman's footrule [12].

Table 1. Relative values for all the metrics applied to the set of images in Fig. 4

Image	Q_1	Q'_1	Q_2	Q_3	Q_4
1	1.00	1.00	1.00	0.91	1.00
2	**0.67**	**0.90**	0.40	1.00	0.81
3	0.10	0.12	0.54	0.81	0.85
4	0.38	0.38	0.79	0.70	0.96

Table 2. Reader A and B vs. metric sorting of images from Fig. 4 in accordance to quality. Top to bottom: best to worse.

A	**B**	Q_1	Q'_1	Q_2	Q_3	Q_4
1	**1**	1	1	1	2	1
2	**2**	2	2	4	1	4
4	**4**	4	4	3	3	3
3	**3**	3	3	2	4	2

Table 3. Evaluation of the no-reference metrics w.r.t. reader grading with the use of the similarity score S in (7). The subindex in S indicates reader A or B.

	Q_1	Q'_1	Q_2	Q_3	Q_4
S_A 1st subset	1.00	1.00	0.50	0.50	0.50
S_B 1st subset	1.00	1.00	0.50	0.50	0.50
S_A all images	0.80	0.80	0.55	0.55	0.40
S_B all images	0.90	0.90	0.60	0.65	0.45

It is basically the l_1-norm of the difference between the reference permutation π_r (from the reader) and the metric Q permutation π_q. Given a set U of m elements (images), a permutation π of this set is defined as a set of indexes mapping to U to produce a particular order of the elements, $\pi : \{1, \cdots, m\} \rightarrow \{1, \cdots, m\}$. The similarity score S of two permutations π_r and π_q is defined as:

$$S = 1 - \frac{\sum_{i=1}^{m} |\pi_r(i) - \pi_q(i)|}{p_{\max}} \ , \tag{7}$$

where $p_{\max}$ is the maximum value of the numerator. It occurs when the permutations are reversed and it can be shown that $p_{\max}$ is equal to $m^2/2$ when m is even and $(m^2 - 1)/2$ when m is odd. Perfect agreement means $S = 1$, and the opposite $S = 0$. The inter-reader agreement for the whole set of 20 images yielded a S score of 0.90. The S scores for the first 4 image subset and the whole set of images are shown in Table 3. The difference in the overall scores for both readers is practically negligible. It is also clear that Q_1 outperforms the other metrics in this experiment with agreement scores of 0.8 and 0.9. The most probable reason is the computation of the metric from normalized space-frequency representation of the image.

6 Conclusions

We have considered four state-of-the-art metrics and their applicability for eye fundus imaging. For fundus auto-focusing, all metrics proved to decrease with respect to the deviation from optimal focus, however strict monotonic decrease was only appreciable for the metric Q_1, based on a directional measure of Rényi entropy. This is most likely due to its robustness to illumination variation. As far as image quality assessment is concerned, we showed that from the considered metrics Q_1 and its modified version Q_1' are the most reliable in terms of agreement with expert assessment, evidenced by average similarity scores of 0.8 and 0.9 with readers A and B, respectively. The results lend strong support to the development of a no-reference metric for fundus imaging based on Rényi entropy.

Acknowledgment. This research has been partly funded by the Spanish Ministerio de Ciencia e Innovación and Fondos FEDER (project DPI2009-08879) and projects TEC2010-09834-E and TEC2010-20307. The first author also thanks the Spanish Ministerio de Educación for an FPU doctoral scholarship.

References

1. Ferzli, R., Karam, L.J.: A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB). IEEE Trans. Image Process 18, 717–728 (2009)
2. Wang, Z., Bovik, A.: Modern image quality assessment. Synthesis Lectures on Image, Video, and Multimedia Processing 2, 1–156 (2006)
3. Gabarda, S., Cristóbal, G.: Blind image quality assessment through anisotropy. J. Opt. Soc. Am. A Opt. Image. Sci. Vis. 24, B42–B51 (2007)
4. Moscaritolo, M., Jampel, H., Knezevich, F., Zeimer, R.: An image based autofocusing algorithm for digital fundus photography. IEEE Trans. Med. Imag. 28, 1703–1707 (2009)
5. Lowell, J., Hunter, A., Steel, D., Basu, A., Ryder, R., Fletcher, E., Kennedy, L.: Optic nerve head segmentation. IEEE Trans. Med. Imag. 23, 256–264 (2004)
6. Zhu, X., Milanfar, P.: Automatic parameter selection for denoising algorithms using a no-reference measure of image content. IEEE Trans. Image Process. 19, 3116–3132 (2010)
7. Qu, Y., Pu, Z., Zhao, H., Zhao, Y.: Comparison of different quality assessment functions in autoregulative illumination intensity algorithms. Optical Engineering 45, 117201 (2006)
8. Abràmoff, M.D., Garvin, M., Sonka, M.: Retinal imaging and image analysis. IEEE Rev. Biomed. Eng. 3, 169–208 (2010)
9. Bartling, H., Wanger, P., Martin, L.: Automated quality evaluation of digital fundus photographs. Acta Ophthalmologica 87, 643–647 (2009)
10. Abràmoff, M.D., Niemeijer, M., Suttorp-Schulten, M.S.A., Viergever, M.A., Russell, S.R., van Ginneken, B.: Evaluation of a system for automatic detection of diabetic retinopathy from color fundus photographs in a large population of patients with diabetes. Diabetes Care 31, 193–198 (2008)
11. Marrugo, A.G., Millán, M.S.: Retinal image analysis: preprocessing and feature extraction. Journal of Physics: Conf Series 274, 12039 (2011)
12. Fagin, R., Kumar, R., Sivakumar, D.: Comparing top k lists. SIAM J. Discrete Math. 17, 134–160 (2003)

Adaptive Medical Image Denoising Using Support Vector Regression

Dinh Hoan Trinh[1], Marie Luong[2], Jean-Marie Rocchisani[3],
Canh Duong Pham[4], and Françoise Dibos[1]

[1] LAGA, Université Paris 13, Avenue Jean-Baptiste Clément, Villetaneuse, France
[2] L2TI, Université Paris 13, Avenue Jean-Baptiste Clément, Villetaneuse, France
[3] Hôpital Avicenne, Bobigny - UFR SMBH, Université Paris 13, Villetaneuse, France
[4] CIID, Vietnam Academy of Science and Technology, Hanoi, Vietnam

Abstract. Medical images are often corrupted by random noise due to various acquisitions, transmission, storage and display devices. Noise can seriously affect the quality of disease diagnosis or treatment. Image denosing is then a required task to ensure the quality of medical image analysis. In this paper, we propose a novel method for reducing some types of common noises in medical images by using a set of given standard images and a well-known machine learning technique namely the Support Vector Regression (SVR). Experimental results are carried out to demonstrate that our method can effectively denoise while preserving small details. A comparison is also performed to demonstrate the outperformance of the proposed technique in terms of both objective and subjective evaluations.

Keywords: Support vector regression, fuzzy c-means, singular value decomposition, medical image, image denoising.

1 Introduction

Denoising is one of the first requirements in the medical image processing. In many denoising methods, the noise is assumed to be normally distributed and additive. However, the nature of noise in medical images such as CT (Computed Tomography), MR (Magnetic Resonance), PET (Positron Emission Tomography) or other modalities can be complex, as these images are generally affected by noise due to various processes from acquisition to display devices. Removing such complex noise is then a difficult task. Unlike conventional denoising methods, a medical image denoising method must remove noise effectively while preserving edges and fine details as much as possible, because subtle details can reveal critical pathological information. This is one of the major challenges in medical image denosing. Generally, image denoising methods can be classified into three main approaches: PDE based approaches [4]-[8], sparcifying transform approaches such as wavelets [9]-[11], and Nonlocal-means based approaches [1]-[3]. The PDE based-approach such as the Total Variation (TV) methods can provide excellent performances in edge preservation and smoothing of flat regions. However, these methods suffer from a staircasing effect in regions with

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 494–502, 2011.

gradual variations [8]. In addition, details and textures can be over-smoothed [1]. The second approach is developed and applied very broadly to medical imaging [11]. However, typical wavelet-based methods can produce significant artifacts because of the structure of the underlying wavelets. Nonlocal means method (NLM) [1] is initially designed to denoise images affected by additive white Gaussian noise with zero-mean and constant-variance. The basis idea is to restore the value of a pixel by computing a weighted average of all pixels in the noisy image. Some extensions of this method for medical image denoising are later introduced [2], [3]. However, it is clear that using a weighted average of all pixels in the noisy image to recover the original image is not guaranteed, especially for non-Gaussian noise. In fact, some of important small details can also be lost. Recently, another solution for denoising relying on learning machine technique is introduced in [12]. Although the results are still far from satisfactory, the idea is interesting.

In medical imaging, we observe the interesting fact that many images can be acquired at approximately the same location. Thus, it is very helpful to use a set of standard (acceptably and proven by experts as noise-free) images to denoise a new noisy image. In this paper, a novel edge/texture-preserving denoising method for medical imaging is proposed. This method can produce better performance than available alternatives and also be used for different types of noise and for any type of medical images. The main idea of the method is to construct an adaptive denoising machine M by using a set of given standard images, the fuzzy c-means clustering technique and the learning method SVR. Machine M is a set of many different SVR functions. Each of them corresponds to a type of noise with a certain noise level around a certain position in the body. Then, denoising can be performed using the SVR functions of machine M.

The rest of this paper is organized as follows. In Section 2, we briefly review the SVR technique. Section 3 describes our proposed algorithm. Our experiments and results are reported in Section 4. Section 5 concludes the article.

2 An Overview of Support Vector Regression

Suppose we are given ℓ observations $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)$ (called training set). Each of observations consists of a vector $\mathbf{x}_i \in \mathbb{R}^d, i = 1, \ldots, \ell$ and the associated "truth" $y_i \in \mathbb{R}$ given by a trusted source. The establishment of the training set will be shown in the next subsection. The goal of the regression is to learn the mapping $\mathbf{x}_i \to y_i$. In general, the training set is not linearly distributed and a conventional linear regression is not sufficient. So, a nonlinear regression function is required for a better estimation. Nonlinear SVR [15] is one of the most well-known techniques to solve this problem and often outperforms other techniques. Its basic idea is to use a mapping function $\phi(\mathbf{x})$ to map the data into a higher dimensional space $\mathcal{H}$ (also called feature space) and then find a linear regression function $y = f(\mathbf{x}) = \langle W, \phi(\mathbf{x})\rangle + b$ according to a new training set $\{(\phi(\mathbf{x}_1), y_1), \ldots, (\phi(\mathbf{x}_\ell), y_\ell)\}$, which represents a nonlinear regression in the original input space. In order to find a linear regression function in the feature space, SVR solves the following optimization problem:

$$\min_{W,b,\xi,\xi^*}\left[\frac{1}{2}\|W\|^2 + C\sum_{i=1}^{\ell}(\xi_i+\xi_i^*)\right] \tag{1}$$

$$\text{subject to}\begin{cases} y_i - \langle W,\phi(\mathbf{x}_i)\rangle - b \leqslant \varepsilon + \xi_i \\ \langle W,\phi(\mathbf{x}_i)\rangle + b - y_i \leqslant \varepsilon + \xi_i^* \\ \xi_i,\xi_i^* \geqslant 0 \end{cases}$$

where $\langle\cdot,\cdot\rangle$ denotes the dot product in $\mathcal{H}$, $C,\varepsilon > 0$ are constants. The primal Lagrangian is as follows:

$$L_P = \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{\ell}(\xi_i+\xi_i^*) - \sum_{i=1}^{\ell}\alpha_i(\varepsilon+\xi_i-y_i+\langle W,\phi(\mathbf{x}_i)\rangle + b)$$

$$-\sum_{i=1}^{\ell}\alpha_i^*(\varepsilon+\xi_i^*+y_i-\langle W,\phi(\mathbf{x}_i)\rangle - b) - \sum_{i=1}^{\ell}(\eta_i\xi_i+\eta_i^*\xi_i^*). \tag{2}$$

It is understood that dual variables in (2) have to satisfy positivity constraints, i.e. $\alpha_i,\alpha_i^*,\eta_i,\eta_i^* \geq 0, \forall i$. From the saddle point condition, we obtain the dual optimization problem of (1):

$$\max_{\alpha_i,\alpha_i^*} -\frac{1}{2}\sum_{i,j=1}^{\ell}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(\mathbf{x}_i,\mathbf{x}_j) - \sum_{i=1}^{\ell}[\varepsilon(\alpha_i+\alpha_i^*) - y_i(\alpha_i-\alpha_i^*)] \tag{3}$$

$$\text{subject to } \sum_{i=1}^{\ell}(\alpha_i-\alpha_i^*) = 0 \text{ and } 0 \leq \alpha_i,\alpha_i^* \leq C, i = 1,\ldots,\ell$$

where $K(\mathbf{x}_i,\mathbf{x}_j) = \langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\rangle$ is called *kernel function*. Solving problem (3) consists in determining the Lagrange multipliers α_i,α_i^*. Let SVs be a set of indexes i where $\alpha_i - \alpha_i^* \neq 0$. Then the SVR function is determined as follows:

$$f(\mathbf{x}) = \sum_{i\in SVs}(\alpha_i-\alpha_i^*)\langle\phi(\mathbf{x}_i),\phi(\mathbf{x})\rangle + b = \sum_{i\in SVs}(\alpha_i-\alpha_i^*)K(\mathbf{x}_i,\mathbf{x}) + b. \tag{4}$$

If $i \in SVs$, $\mathbf{x}_i$ is called a support vector. Commonly, the number of support vectors is much smaller than ℓ. So, SVR is usually faster than traditional regression techniques. From (4), it is clear that the direct mapping $\phi(\mathbf{x})$ is not used. Therefore, there is no need to explicitly obtain $\phi(\mathbf{x})$ as long as we can access the kernel function. Commonly used kernel functions are linear, polynomial, sigmoid, Gaussian. Kernel function should satisfy Mercer's condition (the interested reader is referred to [15] for more details).

3 Proposed Denoising Method

The main idea of the proposed method is to construct a learning machine on a given set of standard images (noise-free images) and then use it for denoising.

Our method is referred as MD (Machine for Denoising) method. Accordingly, the proposed method essentially includes the training phase and the denoising phase. In the training phase, a denoising machine M is constructed and defined by a set of SVR functions, according to the training set which is established from the given standard images and the noisy versions made from these standard images (see subsection 3.1). In order to remove noise adaptively, we use fuzzy c-means (FCM) [14] clustering technique to classify the training set into several groups before determining the SVR functions for M. In the denoising phase, with a noisy image as input, the noise of the image is first determined (type and level of the noise). Then, denoising machine M automatically chooses adaptive SVR functions to estimate the value for each pixel in the image. These phases are detailed in the next subsections.

3.1 Denoising Machine

Denoising machine is designed with many stacks, each of them is designed for one type of image (CT, MR, PET or other modalities). Moreover, each stack includes many substacks. Each substack is designed to image denoising at a certain location in the body such as brain, neck, knee, etc. For each substack, there are several options corresponding to different levels of the noise. Each option has many SVR functions trained from the given standard images.

Assume that we have a database of standard images at approximately the same location, let $N(T, \sigma)$ denote a noise of a type T (Gaussian, Rician, Poisson, etc) with standard deviation σ. In order to establish the training set from a set of standard images, we first add noise N into the standard images. Each observation in the training set is a pair $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i$ is a vector corresponding to a patch of fixed size $(2s + 1) \times (2s + 1)$ and centered at pixel i in the noisy version B established from a standard image A, while y_i is the value of pixel i in image A. By this way, we obtain the training set $G = \{(\mathbf{x}_i, y_i), i = 1, \ldots, \ell\}$ from the standard images. Then, the training set is separated into groups. Each group contains observations such that their patches have some similar features. Here, features are defined according to the following image characteristics: homogenous zone, texture/edge zone and luminance. In order to quantify the luminance of an image patch we can use the average of pixel values in the patch. In the other hand, according to [13], by applying the Singular Value Decomposition (SVD) method to the gradient field of patch $\mathbf{x}_i$, we can quantify its edgeness. For a homogeneous region, there is no dominant direction and all eigenvalues are small. For an oriented edge/texture region, there is a dominant direction and the corresponding eigenvalue is significantly larger than the others.

In summary, for each patch $\mathbf{x}_i$ in the training set we define a characteristic vector $\mathbf{v}_i = (\lambda_1^i - \lambda_2^i, \mu_i) \in \mathbb{R}^2$ with λ_1^i, λ_2^i are singular values of $\mathbf{x}_i$ and μ_i is the mean of pixel values in $\mathbf{x}_i$. The next step concerns the classification of the training set into c groups $(2 \leq c \leq \ell)$, where c can be achieved using histogram of standard images. It can be done by classifying the set of characteristic vectors $\Omega = \{\mathbf{v}_i, i = 1, \ldots, \ell\}$ into c clusters. For an effective classification, a well-known technique, namely fuzzy c-means (FCM) clustering is used. Let $\mathbb{R}^{c\ell}$ denote set of all real $c \times \ell$ matrices. According to [14], FCM algorithm partitions set Ω into c clusters while minimizing the following optimization problem:

$$\min_{U,\nu} J_p(U, \nu) = \sum_{k=1}^{c} \sum_{j=1}^{\ell} u_{kj}^p \|\mathbf{v}_j - \nu_k\|^2 \tag{5}$$

where $1 < p < \infty$, is a constant, $\nu = (\nu_1, \ldots, \nu_c)^T \in \mathbb{R}^{c2}$, $\nu_k \in \mathbb{R}^2$ is the prototype for cluster k and $U = [u_{kj}] \in \mathbb{R}^{c\ell}$ represents a non-degenerate fuzzy c-partition of Ω, the entries of which satisfy:

$$u_{kj} \in [0, 1],\ 1 \leq k \leq c; 1 \leq j \leq \ell,\ \sum_{k=1}^{c} u_{kj} = 1, \forall j, \text{ and } \sum_{j=1}^{\ell} u_{kj} > 0, \forall k. \tag{6}$$

After performing classification, we obtain the training set $G = G_1 \cup G_2 \cup \ldots \cup G_c$ where each group G_k has a characteristic vector $\nu_k \in \mathbb{R}^2$. Consequently, the SVR functions $f_1, f_2, \ldots, f_c$ of an option in the machine M are determined according to the groups $G_1, G_2, \ldots, G_c$, respectively (see section 2).

3.2 Principle of Operation of the Denoising Machine

Let Y be the image to be denoised and $\hat{N}$ an estimation of the noise on Y. The distribution of $\hat{N}$ can be determined according to the opinion of experts. For example, the noises on CT images were found to have a Gaussian probability density function [16]. Noise on MR images has Rician distribution [17] while Poisson noise was found in fluorescent confocal microscopy imaging, X-ray films, PET and SPET. In the proposed method, the first step to denoise Y consists in selecting the suitable option based on the type of image, the position in the body of Y and the estimated standard deviation $\hat{\sigma}$ of the noise $\hat{N}$. Then, denoising is performed by the machine M as follows: for each pixel i in Y, we denote $\mathbf{x}$ in $\mathbb{R}^d$ ($d = (2s + 1)^2$) as a vector corresponding to a patch of size $(2s + 1) \times (2s + 1)$ and centered at i as in the training set (built with the estimated noise of Y; see 3.1). A characteristic vector $\mathbf{v}$ of $\mathbf{x}$ is then computed (see subsection 3.1). In the second step, machine M automatically determines the SVR function that is used for estimating the value of pixel i. SVR function f_k is selected if

$$k = \arg \min_{1 \leq t \leq c} \{\|\mathbf{v} - \nu_t\|^2\}. \tag{7}$$

Finally, true value of pixel i is estimated as follows:

$$f_k(\mathbf{x}) = \sum_{j \in SVs} (\alpha_i - \alpha_j^*) K(\mathbf{x}_j, \mathbf{x}) + b \tag{8}$$

In this paper, the Gaussian function $K(\mathbf{x}_j, \mathbf{x}) = exp(-\|\mathbf{x}_j - \mathbf{x}\|^2)/(2h^2)$ is chosen as kernel function, where h is the decay parameter.

Although the number of the training set is very large, classifying the training set into groups to determine the SVR functions can be done easily. Moreover, the

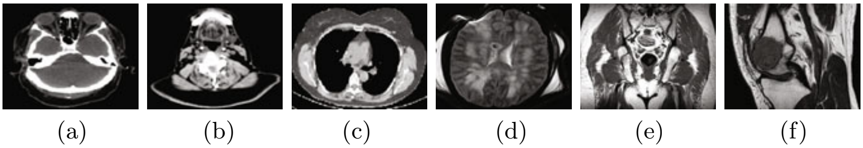


Fig. 1. Test images: CT images of head (a), neck (b), Thorax (c); MR images of head (d), pelvis (e), knee (f)

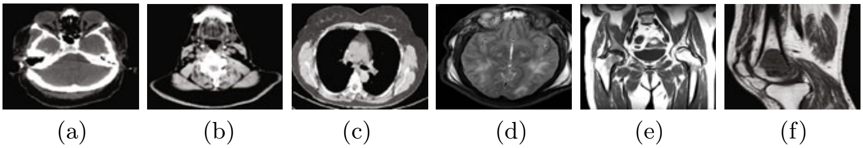


Fig. 2. Standard images of the training set: (a)-(c) CT images, (d)-(f) MR images

estimated value for each pixel in the noisy image only depends on the training group, which includes patches that have similar characteristics as the patch defined for the pixel under considerration. The Gaussian function can be seen as a measurement of similarity between two image patches. Therefore, according to ([8]), $K(\mathbf{x}_j, \mathbf{x})$ may be viewed as weights. The more $\mathbf{x}_j$ is similar to $\mathbf{x}$, the higher is the weight. This shows the adaptiveness of the proposed method.

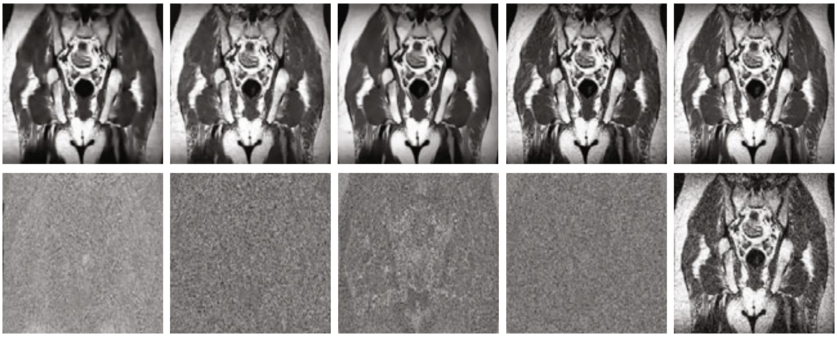
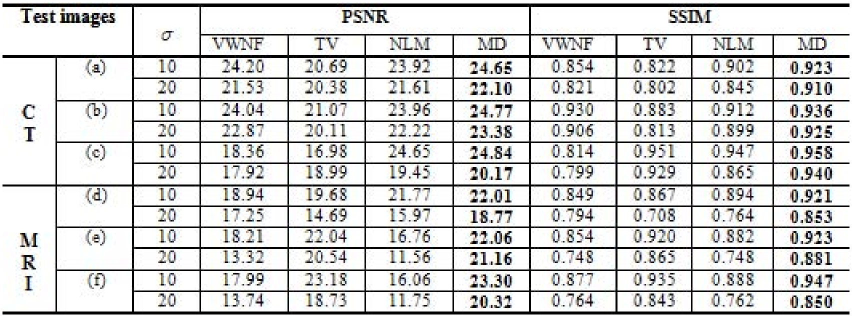


Fig. 3. Result of MR image of the pelvis in Fig. 1(e). From left to right, first row: results of VWNF, TV, NLM, MD and the original image; second row: residual images of VWNF, TV, NLM, MD and the MR image with Rician noise ($\sigma = 20$), respectively.

4 Experimental Results

The proposed MD method is tested on several CT and MR images. We report here some examples of test CT and MR images (Fig. 1). Two quality metrics, namely the PSNR and the SSIM [18] are used to evaluate the performance of our

Table 1. PSNR and SSIM comparison of denoised images



Test images		σ	PSNR				SSIM			
			VWNF	TV	NLM	MD	VWNF	TV	NLM	MD
CT	(a)	10	24.20	20.69	23.92	**24.65**	0.854	0.822	0.902	**0.923**
		20	21.53	20.38	21.61	**22.10**	0.821	0.802	0.845	**0.910**
	(b)	10	24.04	21.07	23.96	**24.77**	0.930	0.883	0.912	**0.936**
		20	22.87	20.11	22.22	**23.38**	0.906	0.813	0.899	**0.925**
	(c)	10	18.36	16.98	24.65	**24.84**	0.814	0.951	0.947	**0.958**
		20	17.92	18.99	19.45	**20.17**	0.799	0.929	0.865	**0.940**
MRI	(d)	10	18.94	19.68	21.77	**22.01**	0.849	0.867	0.894	**0.921**
		20	17.25	14.69	15.97	**18.77**	0.794	0.708	0.764	**0.853**
	(e)	10	18.21	22.04	16.76	**22.06**	0.854	0.920	0.882	**0.923**
		20	13.32	20.54	11.56	**21.16**	0.748	0.865	0.748	**0.881**
	(f)	10	17.99	23.18	16.06	**23.30**	0.877	0.935	0.888	**0.947**
		20	13.74	18.73	11.75	**20.32**	0.764	0.843	0.762	**0.850**

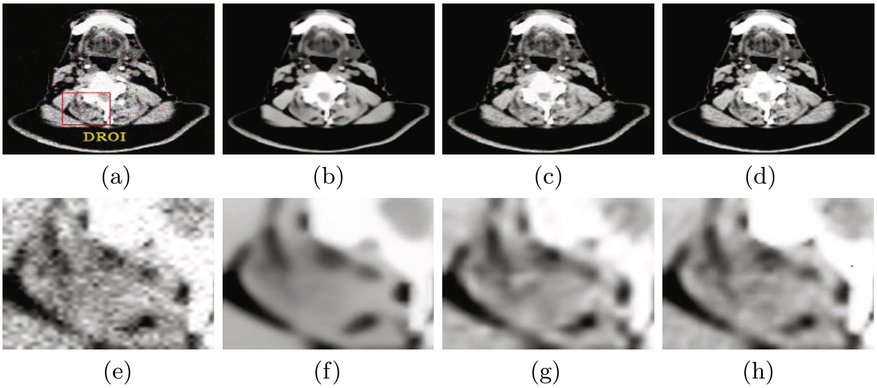


Fig. 4. Results of CT image of the neck in Fig. 1(b): (a) Noisy image with Gaussian noise ($\sigma = 20$), (b) result of NLM, (c) result of MD and (d) Original image. (e)-(h) illustrate zoom-in images of DROI in (a)-(d), respectively.

method in terms of fidelity between the denoised image and the original noise-free image. The SSIM is chosen as it better measures the structure similarity between the recovered image and the reference one, when compared with the PSNR. The original noise-free image is the test image (Fig. 1). Then, the noisy image is obtained by addition of the test image with a noise corresponding to the type of medical image. While CT images are generally corrupted by additive Gaussian noise, MR images are affected by Rician noise. We generate Rician noise by adding two independent Gaussian noises to the real and imaginary part of MR image, respectively. We use the Gaussian noise with zero mean and standard deviation $\sigma = 10$ and 20. The performance of our MD method is compared with three state-of-the-art image denoising methods, namely the TV of Gilboa *et al.* [6], the wavelet based method (VWNF) of Pizurica *et al.* [11] and the NLM. Here, we use the NLM of Buades *et al.* [1] for Gaussian noise on CT images, and the NLM method proposed by Manjón *et al.* [2] for Rician noise on MR images. In our experiments, for each test image, a training set is established by using three standard images. Fig. 2 only illustrates one of those standard images for each example. We use a patch size 5×5. Parameter h of the kernel function in (8) is set to $\hat{\sigma}$ and p in (5) is set to 2. The results obtained for the test MR image in Fig. 1(e) as well as its noisy image affected with Rician noise, are shown in Fig. 3, for visual comparison. We can see that residual image of our method contains nearly no texture or structure while other methods and particularly the NLM contain many edges that have been removed by these denoising methods. Likewise, comparative results between NLM method and the proposed method for the test CT image in Fig. 1(b) (with Gaussian noise) are presented in Fig. 4. Fig. 4(e) - 4(h) illustrate zoom-in images of a desired region of interest (DROI) in Fig. 4(a) - 4(d), respectively. As can be seen, our MD method effectively removes noise while better preserving many subtle details and the textures compared to other methods. In Table 1, it is clear that the proposed method yields a significant PSNR and SSIM gap over the other methods.

5 Conclusion

In this paper, a novel method for medical image denoising is proposed. This method is based on learning machine using the SVR and a given set of standard images. The method can be used for different types of noise, while existing solutions are often designed only for a certain type of noise. Experimental results demonstrated the superior performance of the proposed method over some well known techniques. We believe that with an effective training set, this technique may be quite useful and promising.

References

1. Buades, A., Coll, B., Morel, J.M.: A review of image denoising algorithms, with a new one. Multiscale Modeling and Simulation 4, 490–530 (2005)
2. Manjón, J.V., Carbonell-Caballero, J., Lull, J.J., Garcia-Marti, G., Marti-Bonmati, L., Robles, M.: MRI denoising using nonlocal means. Medical Image Analysis 12(4), 514–523 (2008)
3. Wiest-Daessle, N., Prima, S., Coupe, P., Morrissey, S., Barillot, C.: Rician noise removal by non-local means filtering for low signal-to-noise ratio MRI: Applications to DT-MRI. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part II. LNCS, vol. 5242, pp. 171–179. Springer, Heidelberg (2008)
4. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. In: Physica D, IEEE Conf., vol. 60, pp. 259–268 (1992)
5. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Trans. Pattern Anal. Mach. Intell. 12(7), 629–639 (1990)
6. Gilboa, G., Sochen, N., Zeevi, Y.Y.: Texture preserving variational denoising using adaptive fidelity term. In: Proc. VLSM 2003, Nice, France (October 2003)
7. Dibos, F., Koepfler, G.: Global total variation minimization. SIAM Journal on Num. Anal. 37, 646–664 (2000)
8. Weickert, J.: Anisotropic Diffusion in Image Processing. B. G. Teubner, Stuttgart (1998)

9. Donoho, D.L.: De-noising by soft-thresholding. IEEE, Trans. Inform. Theory 41(5), 613–627 (1995)
10. Chang, S.G., Yu, B., Vetterli, M.: Adaptive wavelet thresholding for image denoising and compression. IEEE Trans. on Image Proc. 9(9), 1532–1546 (2000)
11. Pizurica, A., Philips, W., Lemahieu, I., Acheroy, M.: A versatile wavelet domain noise filtration technique for medical imaging. IEEE Transactions on Medical Imaging 22, 323–331 (2003)
12. Li, D.: Support vector regression based image denoising. Image and Vision Computing 27, 623–627 (2009)
13. Feng, X., Milanfar, P.: Multiscale principal components analysis for image local orientation estimation. Presented at the 36th Asilomar Conf. Signals, Systems and Computers, Pacific Grove, CA (November 2002)
14. Bezdec, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
15. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, N.Y (1995)
16. Lu, H., Li, X., Hsiao, I.T., Liang, Z.: Analytical noise treatment for low-dose CT projection data by penalized weighted least squares smoothing in the K-L domain. In: Proc. SPIE. Medical Imaging, vol. 4682, pp. 146–152 (2002)
17. Gudbjartsson, H., Patz, S.: The rician distribution of noisy MRI data. Magn. Reson. Med. 34, 910–914 (1995)
18. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13(4), 600–612 (2004)

Data-Driven Cortex Segmentation in Reconstructed Fetal MRI by Using Structural Constraints

Benoît Caldairou[1,*], Nicolas Passat[1], Piotr Habas[2], Colin Studholme[2], Mériam Koob[3], Jean-Louis Dietemann[3], and François Rousseau[1]

[1] LSIIT, UMR 7005 CNRS-Université de Strasbourg, France
[2] BICG, University of Washington, Seattle, USA
[3] LINC, UMR 7237 CNRS-Université de Strasbourg, France

Abstract. *In utero* fetal MR images are essential for the diagnosis of abnormal brain development and understanding brain structures maturation. Because of particular properties of these images, such as important partial volume effect and tissue intensity overlaps, few automated segmentation methods have been developed so far compared to the numerous ones existing for the adult brain anatomy. In order to address these issues, we propose a two-step atlas-free cortex segmentation technique including anatomical priors and structural constraints. Experiments performed on a set of 6 *in utero* cases (gestational age from 25 to 32 weeks) and validations by comparison to manual segmentations illustrate the necessity of such constraints for fetal brain image segmentation.

Keywords: Cortex, fetal brain, segmentation, topology.

1 Introduction

The study of *in utero* developing brain by magnetic resonance imaging (MRI) is motivated by the need of understanding the early brain structure maturation [16,12]. A prerequisite is the automated labeling of these structures, which has to be robust to noise, fetal motion artifacts, partial volume effects (PVE), and MRI intensity inhomogeneity.

Other studies focused mainly on premature, noenates and young children. Prastawa *et al.* [14] developed an automated segmentation process of the newborn brain, including estimation of the initial parameters through a graph clustering strategy, intensity inhomogeneity correction and a final refinement focusing on the separation of myelinated and non-myelinated white matter regions. White matter delineation from deep grey matter was also a challenge addressed by Murgasova *et al.* [10] for young children with an atlas-based approach. Another method by Xue *et al.* [17], focusing on cortex segmentation and

[*] The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement no. 207667).

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 503–511, 2011.

reconstruction through a mislabeled partial volume voxel removal strategy was applied to term and preterm neonates.

A first attempt for fetal brain structures segmentation was a semi-automated algorithm based on a region-growing method by Claude *et al.* [4]. Later on, fully automatic techniques were developed. Bach Cuedra *et al.* [1] introduced separated Bayesian segmentation and Markov random field regularization steps, the latter including anatomical priors. Other methods took advantage of motion-corrected and high resolution 3D volumes, computed through reconstruction techniques from *in utero* MR scans [8,15]. Habas *et al.* developed an automatic atlas-based segmentation [7], and a method including anatomical constraints in form of a laminar prior [6]. Gholipour *et al.* [5] performed a volumetric study of the brain based on the segmentation of the pericerebral fluid spaces (PFS) (the part of the cerebrospinal fluid (CSF) located around the cortical area) by using level-sets, connected components, and mathematical morphology filters.

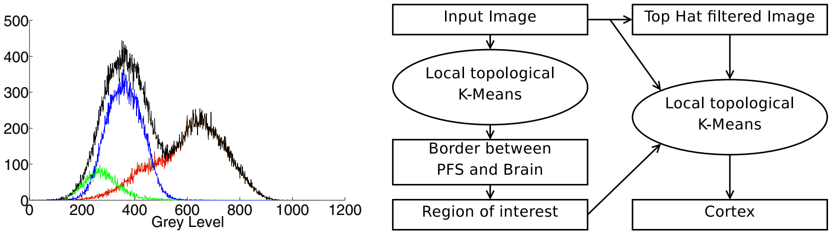
Most of these methods follow an atlas-based approach or specific regularization strategies, including anatomical priors. This illustrates the difficulty to define a data-driven segmentation, because of PVE and important tissue intensity overlaps. Nevertheless, building and using an atlas presents several difficulties such as its registration over the different cases in order to have an accurate segmentation. Moreover, using a specific regularization strategy disconnected from the data illustrates the need of strong structural constraints which can be also used in a data-driven approach.

An atlas-free two-step segmentation is defined. It includes structural constraints based on a topological model [13] in order to deal with PVE, and a morphological filter [11] in order to highlight areas where the cortex will the most likely appear. The first step aims at defining a region of interest including the cortex and the second one aims at segmenting the cortex itself. Experiments are carried out on reconstructed 3D volumes and the probability maps issued from of a non-local fuzzy c-means (NL-FCM) clustering algorithm [3] are used in order to benefit from its robustness to noise.

2 Method

The grey level histogram from fetal MRI (Fig. 1(a)) reveals two peaks corresponding respectively to the brain, including white matter and cortex, and to the CSF. Moreover, an important overlap due to intensity inhomogeneity and partial volume effect is observed. Furthermore, an analysis from a ground truth segmentation reveals that the cortex and white matter intensity distributions are melted into the brain pick, meaning that these structures can not be dissociated by classic clustering algorithms based on intensity features. This leads to hazardous classifications such as white matter between CSF and cortex, which is anatomically wrong.

To cope with the previous problems, a two steps segmentation is defined in order to consider these facts (Fig. 1(b)). Both steps rely on a topological k-means described in Section 2.1. Section 2.2 describes the complete segmentation



(a) (b)

Fig. 1. (a) Grey-level histogram from a fetal brain MRI. Black: intracranial volume, green: cortex, red: white matter and deep grey *nuclei*, blue: CSF. (b) Overall diagram of the segmentation process.

pipeline. The first step aims at separating the intracranial volume into PFS, ventricles and brain. This first segmentation provides a good estimation of the border between the PFS and the brain, which is used to define a region of interest including the cortex. Afterward, the second step is performed in order to retrieve the CSF, the white matter and the cortex.

2.1 Topological K-Means Clustering

A topological model robust to intensity inhomogeneity, relying on three concentric spheres (model already used by [9] for adult brain segmentation) and introducing geometrical constraints for the segmentation process is defined.

Let us consider an image composed of a set of voxels Ω, each voxel $j \in \Omega$ having a given grey-level $\mathbf{y}_j$. Let us suppose that this image has to be segmented into K (≥ 2) clusters. For each cluster k, let S_k be the set of voxel values included into it and ν_k be the centroid of this cluster (which usually corresponds to the mean grey-level value of this class of voxels). Based on these notations, in the k-means approach, the segmentation process of a grey-level image consists of the minimization of an objective function:

$$J_{k\text{-}means} = \sum_{k=0}^{K} \sum_{\mathbf{y}_j \in S_k} \|\mathbf{y}_j - \nu_k\|_2^2.$$

Nevertheless, considering a global centroid (therefore spatially invariant) makes the k-means algorithm sensitive to intensity inhomogeneity occurring in MRI data. In order to tackle this problem without relying on *ad hoc* prior knowledge related to the intensity inhomogeneity, we introduce local intensity centroid values ν_{jk}. These local mean-values are computed in the following way (Fig. 2(a)). An image is divided into several cubical non-overlapping sub-images or regions. Let ν_k^r be the mean value of the kth cluster in an image region r. This region mean value is considered as being located in the center of this region. Let p_r be this position. Afterward, for each considered voxel, a local mean value ν_{jk} is computed by a distance-based interpolation of the nearest region mean-values: $\nu_{jk} = \frac{\sum_r \omega_{jr}\nu_k^r}{\sum_r \omega_{jr}}$, where $\omega_{jr} = 1/d(j, p_r)$ and $d(j, p_r)$ is the Euclidean spatial distance between the position of voxel j and p_r.

The minimization of the k-means objective function is achieved by a border voxel exchange, with respect to the following topological model. Let N_j be the neighborhood of voxel j. Let C_{N_j} be the corresponding set of clusters present in N_j. A considered voxel j switches from cluster k to another candidate cluster k' if it meets the following requirements (Fig. 2(b)):

$$\begin{cases} |C_{N_j}| = 2, \\ \forall\ c \in C_{N_j}, c \neq \text{ background}, \\ \|\mathbf{y}_j - \nu_{jk'}\|_2 < \|\mathbf{y}_j - \nu_{jk}\|_2. \end{cases}$$

The first two requirements guarantee the preservation of the structural constraints. They state that a voxel is eligible for switching from one cluster to another if there are exactly two different clusters in its neighborhood, and if neither of these is the background. The third requirement guarantees that a voxel switch decreases the objective function. Our model is different from the notion of simple points used in topology [13], which implies in particular that labels of connected components are preserved. Broadly speaking, labels of connected components can be broken into several ones or fused as long as the concentric sphere model is respected, which brings a better flexibility to the segmentation process.

In practice, the segmentation is achieved by considering a list of border voxels obtained by a dilation of the current label. Each voxel meeting the third requirement is switched to the considered label. When no switch through the different labels is observed, the centroids are updated and the k-means objective function computed. This process iterates until a local minimum of the objective function is reached.

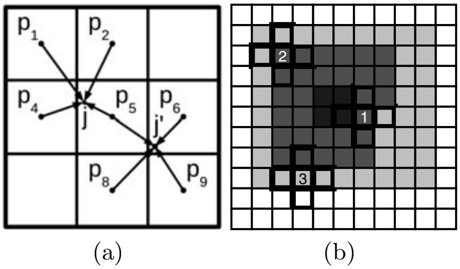


Fig. 2. (a) Intensity inhomogeneity correction. Voxel j mean values depend on mean values from regions 1, 2, 4 and 5 and voxel j' mean values depend on mean values from regions 5, 6, 8 and 9. (b) Topological model. From white to dark grey, labels are 0, 1, 2 and 3, 0 being the background. Voxel 1: not eligible for switching to another label because there are three different labels in its neighborhood and a switch would break the concentric circle model. Voxel 2: eligible to switch to label 1. Voxel 3: not eligible to switch to label 2 because a neighbor is a background label.

2.2 Proposed Segmentation Algorithm

Step 1 - CSF. This step is initialized as follows. The intracranial volume, is divided into three concentric spheres representing the PFS, the brain and the ventricles, thanks to an intracranial distance map (measures the distance from the border of the intracranial volume). Moreover, a two class FCM clustering is performed in order to obtain an accurate initialization of the centroids. The segmentation is then performed by the topological k-means with the grey-level image as input.

Step 2 - Cortex. Due to intensity overlaps between cortex and white matter, additional information is needed in order to achieve the segmentation. Since the fetal cortex is a thin layer between PFS and white matter, a morphological filter is defined in order to highlight image areas where it will most likely appear. Let $I : \Omega \to V$ be a discrete grey-level image. Let φ_B be the morphological closing of I by a structuring element B. The Top Hat Dark Filter T_d is defined as: $T_d(I) = \varphi_B(I) - I$. In other words, this filter highlights small objects of the image that are added by the closing, depending on the choice of the structuring element [11].

A region of interest is defined from the border between the PFS and the brain. A band around this border, including CSF and brain is defined thanks to a distance map computed from the PFS segmentation (measures a distance from the PFS border with the rest of the brain). This band is divided into three sub-bands being the initialization for CSF, cortex and white matter. Moreover, this initialization is corrected by removing any voxel belonging to the ventricles.

The segmentation is performed by using a vector image composed by the original image and the top-hat filtered image, instead of the original grey-level values alone, as the input of the topological segmentation presented in Section 2.1. Consequently, each cluster is characterized by a centroid vector composed of its grey-level mean-value and its top-hat-filtered image mean value, allowing a better discrimination of the cortex. During this process, a maximum cortical thickness of 4 millimeters is imposed in order to cope with improbable extensions.

In order to improve the segmentation, one can use probability maps computed from a non local FCM algorithm [3] as a post-processing. This method introduces a regularization based on a non-local framework [2], aiming at correcting artifacts due to noise, by taking advantage of the redundancy present in images. Broadly speaking, a small neighborhood around a voxel, called a patch, may match patches around other voxels within the same scene, selecting the most accurate voxels to perform the regularization. This post-processing step is run on the same border voxel exchange basis than the topological k-means algorithm, unless a voxel wil lswitch if the probability it belongs to the destination label is higher than the current one.

3 Experiments and Results

3.1 Material and Experimental Settings

Experiments are performed on a set of six patients. Gestational ages (GA) range from 27 to 32 weeks. For each of them, a set of three T2-weighted MR images (axial, coronal and sagittal) are acquired from a 1.5T scanner (Magnetom Avanto, Siemens, Germany Erlangen) using single shot fast spin echo sequences (TR 3190 ms, TE 139-147 ms). Since these images have anisotropic voxel sizes (from $0.742 \times 0.742 \times 3.45$ to $0.742 \times 0.742 \times 4.6$ mm) and may present motion artifacts, a reconstruction process [15] is applied in order to obtain high resolution images.

Reconstructed images have the following dimensions: $256 \times 256 \times 88$ to $256 \times 256 \times 117$ and voxel dimensions are: $0.742 \times 0.742 \times 0.742$ mm. A 3D 6-neighborhood is used to run the topological model. Empirically, a $5 \times 5 \times 5$ structuring element is chosen to perform the Top Hat filter.

For the PFS segmentation, the model is initialized as follows. On the border with the background, a 1 voxel thin layer is set as PFS. Then, the voxels being less than 80% of the maximum intracranial distance are set as brain and the remaining ones are set as ventricles. This guarantees that the ventricles initial cluster will not include any PFS voxels.

Regarding the cortex segmentation, the model was initialized as follows. The first two-millimeters layer is set as CSF, the next 5 as cortex and the last 2 as white matter. These values were chosen according to tissues anatomical characteristics.

Concerning the parameters of the non-local FCM algorithm, the size of the research area is $11 \times 11 \times 11$ and the size of the patches is $3 \times 3 \times 3$. The computation times are about 20 minutes for the extraction of the CSF and 15 minutes for the cortex segmentation, mostly due to the exploration of voxels neighborhoods.

3.2 Validation

Each reconstructed image has been manually segmented. The validation consists of the computation of the dice similarity coefficient (DSC) between the manual

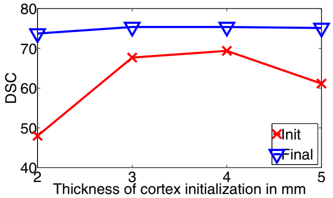
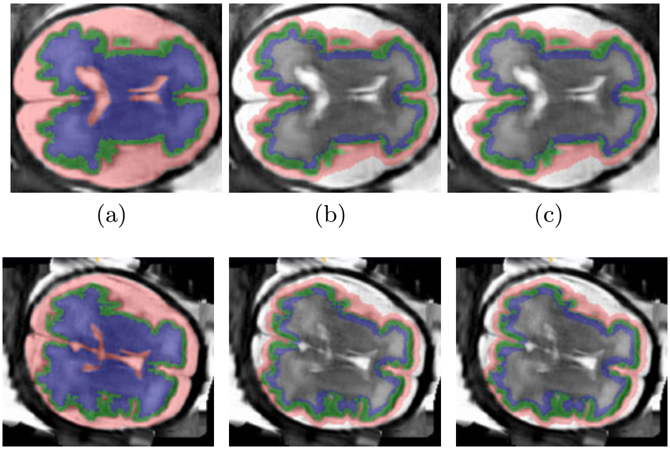


Fig. 3. Average DSC comparison between cortex thickness initialization (red) and final segmentation (blue)

Table 1. Dice similarity coefficient (DSC) between manual and automated segmentations with and without non-local FCM post-processing

Case (GA)	1 (28)	2 (30)	3 (-)	4 (32)	5 (27)	6 (30)
DSC	72.42	73.03	77.57	74.51	76.60	76.61
DSC with regularization	71.02	71.37	77.76	71.86	75.78	75.23


(a) (b) (c)

(d) (e) (f)

Fig. 4. Cortex extraction. (a,d): ground truth, (b,e): segmentation without non-local FCM regularization, (c,f): segmentation with non-local FCM regularization. Red: CSF, green: cortex, blue: white matter and deep grey *nuclei*.

and the automated segmentation of the cortex. Let TP be the amount of true positives (number of detected cortex voxels), FP the amount of false positives (number of voxels incorrectly classified as cortex) and FN the amount of false negatives (number of undetected cortex voxels). The dice coefficient is given by: $DSC = 2 \times TP/(2 \times TP + FN + FP)$.

Table 1 presents DSC for each case. Both regularized and non-regularized results are presented. Fig. 3 illustrates the algorithm robustness to initialization by setting a 2 to 5 millimeters initial thickness to the cortex.

A visual insight of the segmentation (Fig. 4) underlines the accuracy of the method, even though a slight under-segmentation can be observed in some areas. Moreover, even though regularization accentuates the under-segmentation, it can be observed that it brings a noise correction and smoother borders between the different tissues.

Other studies about fetal brain segmentations highlighted results about the cortex segmentation. Bach Cuedra *et al.* [1] showed DSC values around 65 % with a two steps segmentation separating LCR into PFS and ICSF and applying a specific regularization step. Habas *et al.* [7] achieved performance around 82

% with an atlas based approach. Results presented here underline the usefulness of structural constraints for fetal tissue segmentation, if no atlas is available.

4 Conclusion

A topological based clustering method has been proposed for the segmentation of the cortex in fetal brain MR images, which takes advantages of anatomical knowledge. The validation performed on T2-weighted images illustrates the usefulness of such structural constraints in an atlas-free approach of fetal brain segmentation.

Further work will focus on the improvement of the segmentation method, such as a better integration of the regularization step into the process, its validation on additional cases and the segmentation of other tissues and structures of the fetal brain.

References

1. Bach Cuedra, M., Schaer, M., André, A., Guibaud, L., Eliez, S., Thiran, J.-P.: Brain tissue segmentation of fetal MR images. In: Image Analysis for the Developing Brain, Workshop in MICCAI (2009)
2. Buades, A., Coll, B., Morel, J.M.: A review of image denoising algorithms, with a new one. Multiscale Modeling & Simulation 4, 490–530 (2005)
3. Caldairou, B., Passat, N., Habas, P.A., Studholme, C., Rousseau, F.: A non-local fuzzy segmentation method: Application to brain MRI. Pattern Recognition (in press), doi:10.1016/j.patcog.2010.06.006
4. Claude, I., Daire, J.-L., Sebag, G.: Fetal brain MRI, segmentation and biometric analysis of the posterior fossa. IEEE Transactions on Biomedical Engineering 51, 617–626 (2004)
5. Gholipour, A., Estroff, J., Barnewolt, C., Connolly, S., Warfield, S.: Fetal brain volumetry through MRI volumetric reconstruction and segmentation. International Journal of Computer Assisted Radiology and Surgery (in press), doi: 10.1007/s11548-010-0512-x
6. Habas, P.A., Kim, K., Chandramohan, D., Rousseau, F., Glenn, O.A., Studholme, C.: Statistical model of laminar structure for atlas-based segmentation of the fetal brain from in-utero MR images. In: SPIE, vol. 7259, pp. 17–24 (2009)
7. Habas, P.A., Kim, K., Rousseau, F., Glenn, O.A., Barkovich, A.J., Studholme, C.: Atlas-based segmentation of developing tissues in the human brain with quantitative validation in young fetuses. Human Brain Mapping 31, 1348–1358 (2010)
8. Kim, K., Habas, P.A., Rousseau, F., Glenn, O.A., Barkovich, A.J., Studholme, C.: Intersection based motion correction of multislice MRI for 3-D in utero fetal brain image formation. IEEE Transactions on Medical Imaging 29, 146–158 (2010)
9. Miri, S., Passat, N., Armspach, J.-P.: Topology-preserving discrete deformable model: application to multi-segmentation of brain MRI. In: Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D. (eds.) ICISP 2008 2008. LNCS, vol. 5099, pp. 67–75. Springer, Heidelberg (2008)
10. Murgasova, M., Dyet, L., Edwards, D., Rutherford, M., Hajnal, J., Rueckert, D.: Segmentation of brain MRI in young children. Academic Radiology 14, 1350–1366 (2007)

11. Najman, L., Talbot, H.: Mathematical morphology: from theory to applications. ISTE / J. Wiley & Sons (2010)
12. Perkins, L., Hughes, E., Srinivasan, L., Allsop, J., Glover, A., Kumar, S., Fisk, N., Rutherford, M.: Exploring cortical subplate evolution using magnetic resonance imaging of the fetal brain. Developmental Neuroscience 30, 211–220 (2008)
13. Pham, D.L., Bazin, P.-L., Prince, J.L.: Digital topology in brain imaging. IEEE Signal Processing Magazine 27, 51–59 (2010)
14. Prastawa, M., Gilmore, J.H., Lin, W., Gerig, G.: Automatic segmentation of MR images of the developing newborn brain. Medical Image Analysis 9, 457–466 (2005)
15. Rousseau, F., Glenn, O., Iordanova, B., Rodriguez-Carranza, C., Vigneron, D., Barkovich, J., Studholme, C.: Registration-based approach for reconstruction of high-resolution in utero fetal MR brain images. Academic Radiology 13, 1072–1081 (2006)
16. Rutherford, M., Jiang, S., Allsop, J., Perkins, L., Srinivasan, L., Hayat, T., Kumar, S., Hajnal, J.: MR imaging methods for assessing fetal brain development. Developmental Neurobiology 68, 700–711 (2008)
17. Xue, H., Srinivasan, L., Jiang, S., Rutherford, M., Edwards, A.D., Rueckert, D., Hajnal, J.V.: Automatic segmentation and reconstruction of the cortex from neonatal MRI. NeuroImage 38, 461–477 (2007)

Evaluation of Facial Reconstructive Surgery on Patients with Facial Palsy Using Optical Strain

Matthew Shreve[1], Neeha Jain[1], Dmitry Goldgof[1], Sudeep Sarkar[1],
Walter Kropatsch[2,*], Chieh-Han John Tzou[3], and Manfred Frey[3]

[1] University of South Florida, Department of Computer Science and Engineering,
Tampa Florida
{mshreve,neehajain,goldgof,ssarkar}@cse.usf.edu
[2] Vienna University of Technology, Vienna Austria
krw@prip.tuwien.ac.at
[3] Medical University of Vienna, Division of Plastic and Reconstructive Surgery,
Vienna Austria
{chieh-han.tzou,manfred.frey}@meduniwien.ac.at

Abstract. We explore marker-less tracking methods for the purpose of evaluating the efficacy of facial re-constructive surgery on patients with facial palsies. After experimenting with several optical flow methods, we choose an approach that results in less than 2 pixels in tracking error for 15 markers tracked on the face. A novel method is presented that utilizes the non-rigid deformation observed on facial skin tissue to visualize the severity of facial paralysis. Results are given on a dataset that contains three videos of an individual recorded using a standard definition camera both before and after undergoing facial reconstructive surgery over a period of three years.

Keywords: Optical Flow, Optical Strain, Facial Palsy, Facial Reconstructive Surgery.

1 Introduction

Accurately estimating and quantifying the extent of facial paralysis in patients with facial palsy without the need of manually applied markers would be a benefit to patients, researchers, and the medical community at large. In this paper, we propose methods that can be used to measure the severity of facial paralysis using non-invasive tracking methods and motion analysis tools.

The experimental flow is as follows: first, a patient is recorded in front of a video camera mirror system [2] and is asked to perform several standardized expressions multiple times (ex., lifting of eyebrows, smile, close eyes, frown, whistle) [4]. Next, a dense optical flow method is used that tracks all points (pixels) of the face over the entire length of the expressions. These optical flow vectors are then used to calculate optical strain, a feature that is used for two purposes:

[*] W.K. acknowledges the support of the Austrian Science Fund (FWF), grant no. P20134-N13.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 512–519, 2011.

(i) the magnitude of optical strain is utilized in order to detect key moments of an expression (contract, peak, compression) [6]. Finding the these moments in an expression allows strain maps to be calculated at the maximal point of facial deformation, so a valid comparison can be done over time; (ii) strain maps are used to represent and quantize the deformation of the soft-skin tissue on the face, which is directly correlated with expansion and contraction of underlying facial muscles that have been surgically altered.

Evaluating the efficacy of facial reconstructive surgery has been the main goal of Frey *et al.* [2]. In their experimental setup, a patient is asked to sit between two angled mirrors ($\sim 90\,^\circ$). Hand placed markers are applied to the face and are tracked in 3-D as the patient performs expressions. In their setup, the process of applying markers and tracking them takes roughly five hours. In this paper, we use a video dataset from their collection and hope to expand on their initial work firstly by eliminating the need for markers, thus significantly reducing the time needed for data acquisition. Secondly, we suggest a method that provides a denser correspondence and a more detailed visual representation and quantization.

2 Background

When calculating optical strain there are typically two main approaches: either (i) integrate the strain definition into the optical flow equations, or (ii) derive strain directly from the flow vectors. The first approach requires the calculation of high order derivatives, hence is sensitive to image noise. The second approach allows us to post-process the flow vectors before calculating strain, possibly reducing the effects of any errors incurred during the optical flow estimation. We use the second approach in this paper.

2.1 Optical Flow

Optical flow is an established motion estimation technique that is based on the brightness conservation principle [1]. In general, it assumes that the intesntity at a point remains constant over a pair of frames, and that the pixel displacement relatively smooth within a small image region. It is typically represented by the following equation:

$$(\nabla I)^T \mathbf{p} + I_t = 0 \qquad (1)$$

where $I(x, y, t)$ represents the temporal image intensity function at point x and y at time t, and ∇I represents the spatial and temporal gradient. The horizontal and vertical motion vectors are represented by $\mathbf{p} = [p = dx/dt, q = dy/dt]^T$s.

Since large intervals over a single expression can often cause failure in tracking (due to the smoothness constraint), we implemented a vector linking (or stitching) process that combines small, local pairs of small intervals (1-3 frames) into larger pairs to expand over the entire sequence of frames. In section 3.1, we discuss three seperate implementations of optical flow.

2.2 Optical Strain

The displacement of any deformable object projected on a 2-D plane can be expressed by a vector $\mathbf{u} = [u, v]^T$. Assuming a small enough motion , a finite strain tensor can be defined:

$$\varepsilon = \frac{1}{2}[\nabla \mathbf{u} + (\nabla \mathbf{u})^T], \tag{2}$$

which can be expanded to:

$$\varepsilon = \begin{bmatrix} \varepsilon_{xx} = \frac{\partial u}{\partial x} & \varepsilon_{xy} = \frac{1}{2}(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}) \\ \varepsilon_{yx} = \frac{1}{2}(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}) & \varepsilon_{yy} = \frac{\partial v}{\partial y} \end{bmatrix} \tag{3}$$

where (ε_{xx},ε_{yy}) are normal strain components and (ε_{xy},ε_{yx}) are shear strain components.

Since (u,v) are displacement vectors that over a continuous space, we approximate the strain components using the optical flow data (p,q):

$$p = \frac{\delta x}{\delta t} \approx \frac{\Delta x}{\Delta t} = \frac{u}{\Delta t}, u = p\Delta t, \tag{4}$$

$$q = \frac{\delta y}{\delta t} \approx \frac{\Delta y}{\Delta t} = \frac{v}{\Delta t}, v = q\Delta t \tag{5}$$

where Δt is the change in time between two image frames. Setting Δt to a fixed interval length, we can estimate the partial derivatives of (4) and (5):

$$\frac{\partial u}{\partial x} = \frac{\partial p}{\partial x}\Delta t, \frac{\partial u}{\partial y} = \frac{\partial p}{\partial y}\Delta t, \tag{6}$$

$$\frac{\partial v}{\partial x} = \frac{\partial q}{\partial x}\Delta t, \frac{\partial v}{\partial y} = \frac{\partial q}{\partial y}\Delta t, \tag{7}$$

The second order derivatives are calculated using the central difference method. Hence,

$$\frac{\partial u}{\partial x} = \frac{u(x + \Delta x) - u(x - \Delta x)}{2\Delta x} \approx \frac{p(x + \Delta x) - p(x - \Delta x)}{2\Delta x} \tag{8}$$

$$\frac{\partial v}{\partial y} = \frac{v(y + \Delta y) - v(y - \Delta y)}{2\Delta y} \approx \frac{q(y + \Delta y) - q(y - \Delta y)}{2\Delta y} \tag{9}$$

where (Δx,Δy) $\approx$ 2-3 pixels.

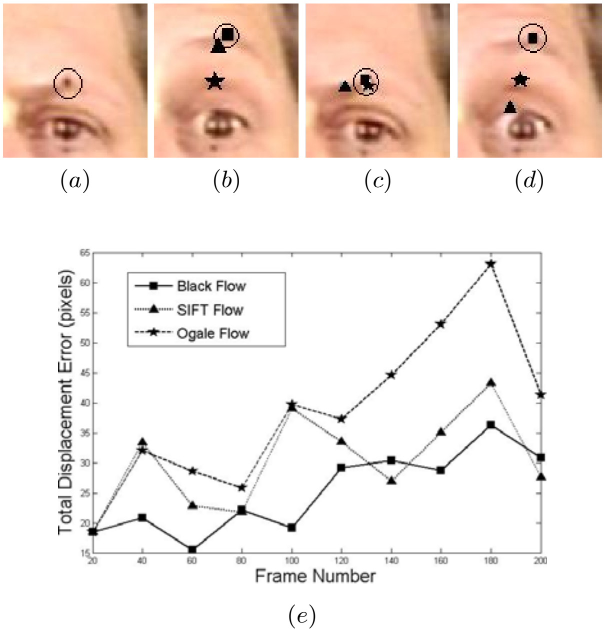
Finally, each of these values corresponding to low and large elastic moduli are summed to generate the strain magnitude. Each value can also be normalized to 0-255 for a visual representation (strain map).

3 Experiment

In this section, we explore several potential uses of optical flow and optical strain for the marker-less tracking and visualization of expressions for patients with facial palsies. Our dataset consists of three videos from the Medical University of Vienna. Each video corresponds to a different year of the patient undergoing facial reconstructive surgery. The first video records the patient before the surgery (1998), and the next two videos (1999 and 2000) were recorded post surgeries. For each video, there are roughly 30 expression made. Expressions include raising the eyebrows, smiling, smiling and closing eyes, bunching lips together, and frowning.

3.1 Optical Flow and Tracking

In this paper, the primary purpose of optical flow is to calculate a dense correspondence between pixels over video sequences that contain expressions, a task that is important for the accurate calculation of strain maps. Hence, we explored several implementations of optical flow, including Ogale flow [5], SIFT flow [3], and Black Flow [1]. To determine the best implementation choice, we inspected



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

(e)

Fig. 1. Example tracking results of point given in circle (a). In (b) - (d), results for black flow (square) Ogale flow (triangle) and SIFT flow (star) at during two 'raise eyebrows' expressions (frame numbers 30, 120, 150). In (e) the actual error for all 15 points (see Fig. 2) is shown for every 20 frames.

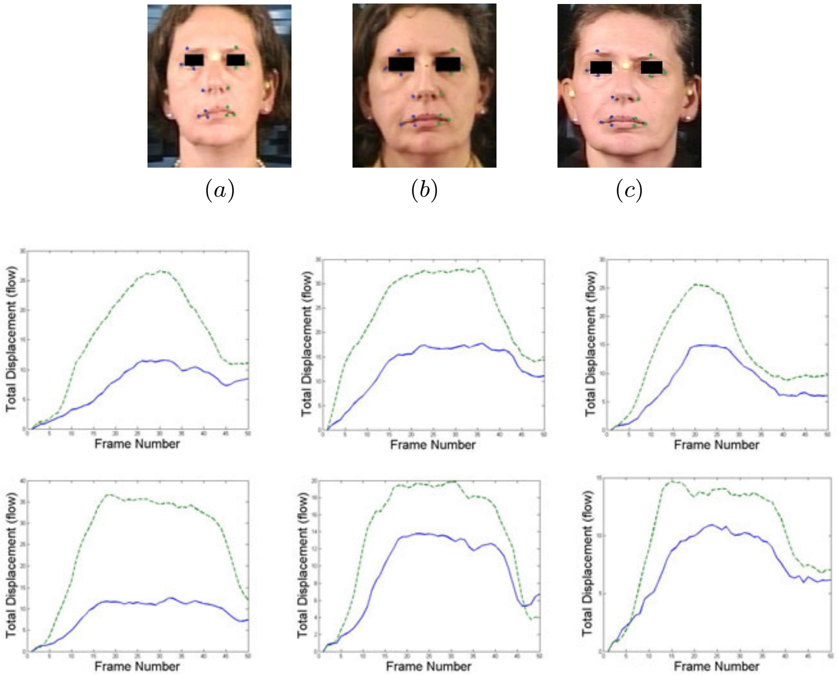


Fig. 2. Comparison of total flow displacement values between the left (solid blue line) and right (dotted green line) sides of face, after re-constructive surgery over 3 years (using Black flow). The images (a), (b) and (c) are the starting frames from each video and show the tracked points. The first row of graphs corresponds to the raised eyebrows expression and the second row corresponds to the smile expression.

the tracking performance over several expressions at specific points. The points selected were the physical markers placed on the face, since these areas have texture information which aids in optical flow estimation. An example sequence containing the raised eyebrows expression can be seen in Fig. 1. This figure also shows the total summed error that was calculated for the same expression every 20 frames, at all fifteen points given in Fig. 2.

To further analyze the tracking results of each flow algorithm on this sequence, we calculated the average error (see Table 1) for all fifteen points and also for a subset of three select points near the right eye (see Fig. 2) where there was large eye / eyebrow motion. A few observations were made: Ogale flow occasionally showed sporadic tracking by jumping several pixels off and then back again. Overall, it resulted in average error rates of 2 pixels (for fifteen points) and 4.3 pixels (for three points). On the other hand, SIFT flow performed poorly even with small non-rigid movements of the eyebrow, since such local motion was dominated back the lack of motion in surrounding regions. It had average error rates of 2.5 pixels (all points) and 6.1 pixels (three points). Black flow performed

Table 1. Average error (in pixels) for all 15 points tracked on face and a subset of 3 points that have relatively large motion

Flow Type	All Points	Three Points
Black Flow	1.67	2.58
SIFT Flow	2.01	4.35
Ogale FLow	2.55	6.14

the best of all three and led to the most consistent results, with average error rates of 1.6 (all points) and 2.5 (three points).

Next, we show the tracking results using black flow for two expressions (raised eyebrows, smile) for each year. For this, points on each side of the face were

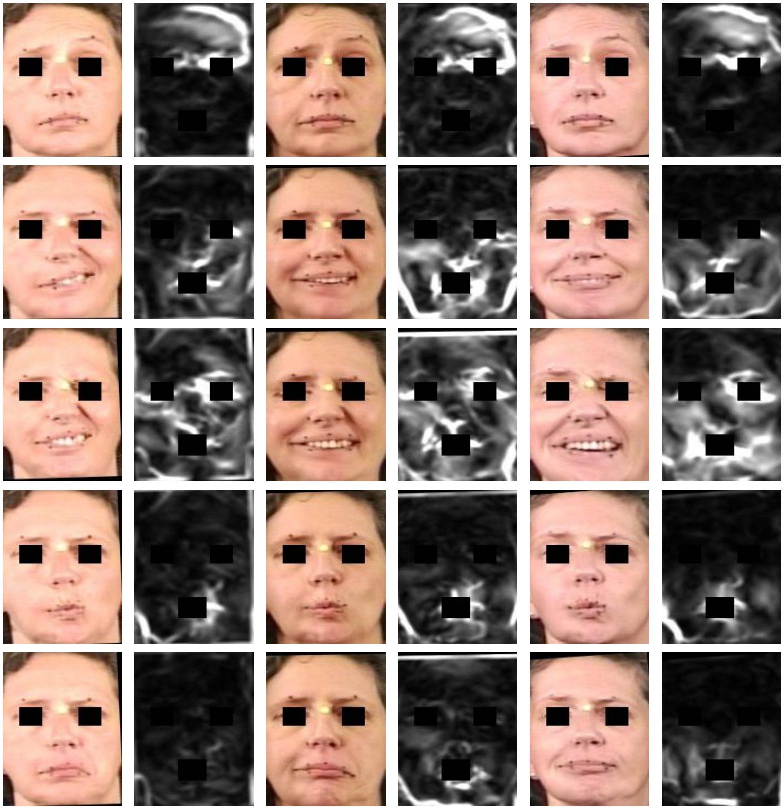


Fig. 3. Optical strain maps for five expressions, over three years. Strain maps were generated between the start and peak of each expression. Intensity values correspond to amount of deformation observed.

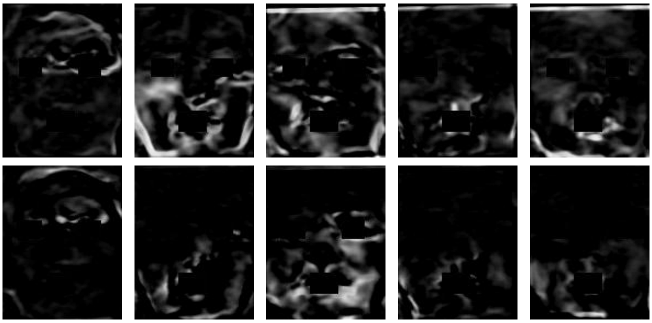


Fig. 4. Quantization using strain map difference between years recorded, post surgery. Each column contains one of the five expressions and shows the difference between strain map rows in Fig 3. The first row shows the change from 1998 and 1999, and the second row shows the change from 1999 to 2000.

tracked over two expressions and the displacements were summed to generate total summed displacement. As expected, for both expressions, the difference between the total displacement observed for each half of the face is largely reduced between 1998 and 1999, and even more between the 1998 and 2000 (see Fig. 2). This indicates that optical flow is successfully capturing the motion caused from the facial expressions. Next, we will use these flow vectors to calculate strain magnitude and optical strain maps.

3.2 Optical Strain Maps

Since strain maps represent the non-rigid deformation observed on the face during an expression, it is important that we capture the peak of the expression. We automatically get this frame using an expression spotting algorithm [6]. In summary, the algorithm utilizes the strain magnitude calculated over the entire video sequence, and correlates spatio-temporal regions that contain high strain values as segments containing expressions. It is particularly robust to expressions that occur in small regions or one side of the face, making it ideal for patients with facial palsy. The algorithm returns the frame number in a expression sequence that has the highest summed strain magnitude. These frames are then used for calculating final strain maps. Fig. 3 shows the strain maps calculated for all five expressions, over all three years. It is important to note here that eye regions and mouth regions have been masked due to common flow failure in these regions, due to self-occlusion (eyelids, inside mouth). For areas outside of the masked regions, large intensity values correspond to regions of the patients soft-skin tissue that have deformed significantly due to muscular contraction.

Quantization using Strain Difference. Subtracting two strain maps from different years but the same expression allows us to gain a representation of the change in deformation, or the change in active regions of the face. As can be observed in Fig. 4, the strain maps showing the difference between the years 1998 and 1999 suggest a large amount of improvement (first row), while the gain between 1999 and 2000 (second row) appears to be less.

4 Conclusions

In this paper, we explore the use of marker-less tracking methods for the purpose of evaluating the improvement gained from facial re-constructive surgery on patients with facial palsies. We have explored several tracking methods that allow us to create the dense correspondence necessary for strain map calculation and have concluded the Black flow leads to the most consistent and reliable results, with less than 2 pixel average tracking error. Using these optical flow fields, we have proposed a method that quantizes the non-rigid deformation observed on facial skin tissue into strain maps. Strain maps can then then be used to highlight the (a)symmetries between each side of the face, while also providing a useful measure of the changes at each point on the face over time, thus potentially allowing surgeons to quickly evaluate the efficacy of facial reconstrutive surgeries.

References

1. Black, M.J., Anandan, P.: The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. In: Computer Vision and Image Understanding, New York, NY, USA, vol. 63, pp. 75–104. Elsevier Science Inc., Amsterdam (1996)
2. Frey, M., Giovanoli, P., Slameczka, M., Stussi, E.: Three-dimensional Video Analysis of Facial Movements: A New Method to Assess the Quantity and Quality of the Smile. Journal of Otology and Neurotology 23, 1531–7129 (1999)
3. Liu, C., Yuen, J., Torralba, A., Sivic, J., Freeman, W.T.: Sift flow: dense correspondence across different scenes. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 28–42. Springer, Heidelberg (2008)
4. Frey, M., Jenny, A., Giovanoli, P., Stussi, E.: Development of a new documentation system for facial movements as a basis for the international registry for neuromuscular reconstruction in the face. Plastic and Reconstructive Surgery 93, 1334–1350 (1994)
5. Ogale, A.S., Aloimonos, Y.: Shape and the stereo correspondence problem. International Journal on Computer Vision 65, 147–162 (2005)
6. Shreve, M., Godavarthy, S., Goldgof, D., Sarkar, S.: Macro- and micro-expression spotting in long videos using spatio-temporal strain. In: The Ninth IEEE International Conference on Automatic Face and Gesture Recognition, FG 2011 (2011)

Inferring the Performance of Medical Imaging Algorithms

Aura Hernàndez-Sabaté[1], Debora Gil[1], David Roche[2],
Monica M.S. Matsumoto[3], and Sergio S. Furuie[3]

[1] Computer Science Department and Computer Vision Center - UAB, Bellaterra, Spain
[2] Laboratory of Systems Pharmacology and Bioinformatics, UAB, Bellaterra, Spain
[3] Faculdade de Medicina and Escola Politécnica da USP, São Paulo, Brazil
{aura,debora}@cvc.uab.es

Abstract. Evaluation of the performance and limitations of medical imaging algorithms is essential to estimate their impact in social, economic or clinical aspects. However, validation of medical imaging techniques is a challenging task due to the variety of imaging and clinical problems involved, as well as, the difficulties for systematically extracting a reliable solely ground truth. Although specific validation protocols are reported in any medical imaging paper, there are still two major concerns: definition of standardized methodologies transversal to all problems and generalization of conclusions to the whole clinical data set.

We claim that both issues would be fully solved if we had a statistical model relating ground truth and the output of computational imaging techniques. Such a statistical model could conclude to what extent the algorithm behaves like the ground truth from the analysis of a sampling of the validation data set. We present a statistical inference framework reporting the agreement and describing the relationship of two quantities. We show its transversality by applying it to validation of two different tasks: contour segmentation and landmark correspondence.

Keywords: Validation, Statistical Inference, Medical Imaging Algorithms.

1 Introduction

Researchers agree that validation of medical imaging algorithms is essential for supporting their validity and applicability in clinical practice [1]. Although validation is addressed in any medical imaging paper, there is no consensus in the statistical and mathematical tools required for standardized quantitative analysis [2,3,1,4]. Given the diversity of imaging tasks and final clinical applications, techniques are prone to be validated using specific protocols, not easily extendable to a unifying general framework [1].

A validation protocol should face two main challenges: extracting ground truth (GT) and defining a metric quantifying differences between GT and the algorithm output (AO). A main difficulty in medical imaging is that GT might not be always available or might vary across observers [5]. The first case is common in image registration tasks, since the deformation matching images might not be easily extracted from in vivo cases. Current solutions, base validation on either synthetic experiments or correspondence of

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 520–528, 2011.

anatomical landmarks [6]. The realism of synthetic databases might be too low for generalization of conclusions to clinical data [5]. It follows that, in real data, a verification based on structures (landmarks) correspondence is usually required. Variability in GT typically arises in segmentation tasks, due to discrepancies across manual tracers. This implies that an analysis of automated errors might not reflect, by its own, the true accuracy of segmentations, since variations might be caused by a significant difference among expert models. A standard solution [7] is comparing automated errors to the variability among different manual segmentations. Concerning comparison between GT and automatic computations, several metrics can be considered. For landmark correspondence, the difference in positions is the accepted goodness measure [6], while for contour segmentation [8,3] differences can be measured by means of area overlap or distances between contours. The counterpart of these metrics is that they assess complementary quality scores and, thus, several quality measures need to be considered.

Two major concerns still remain: 1) defining the subset of scores best reflecting accuracy for clinical application and 2) whether the results of validation tests are generalizable to all clinical data. We claim that a validation protocol assessing to what extent an image processing algorithm can substitute the manual interaction would address both issues. In this context, validation should report the agreement between AO and GT, as well as, a model describing the relation between both quantities.

Agreement between observers can be assessed by means of Bland-Altman plots or regression analysis. Bland-Altman [9] measures this agreement by analyzing the variability of their differences. In the case of disagreement, Bland-Altman fails to either describe or report the degree of disagreement [10]. Regression analysis provides a (linear) statistical model of the relation between two quantities. Existing techniques usually only report regression coefficients (slope and intercept) and correlation. Given that correlation only reports the degree of linear dependence between both quantities, a high correlation does not imply that the variables agree [10]. In order to explore agreement, one should consider the slope and intercept of the regression model, since they describe the relation between the two variables. However, even in the case of a perfect relation (identity), the regression coefficients alone are not sufficient to ensure that the quantities can be swapped. The slope and intercept describe the behavior of the specific sample we are analyzing, but they do not allow to generalize conclusions to the whole population. The only way to obtain generalizable conclusions is by means of statistical inference.

We present a statistical inference framework for assessing how well two methodologies performing the same task behave equally and can replace one each other. We define a regression model for predicting the performance of an image processing algorithm in clinical data from a subset of validated samples. Our model is applied to two main tasks involved in medical image processing: detection (registration) and segmentation of anatomical structures. Experiments on vessel wall segmentation show the correlation between our model and standard metrics. Meanwhile, experiments on cardiac-phase detection illustrate its versatility for assessing difficult tasks.

2 Inference Model

We note by GT the ground truth we want to substitute and AO, the algorithm output. Their nature is prone to vary depending on the particular problem we are facing:

1. **Segmentation.** Image segmentations produce a (continuous) contour enclosing the area of interest. Therefore, $GT = GT(t) = (GT_1(t), GT_2(t))$, $AO = AO(t) = (AO_1(t), AO_2(t))$ are curves parameterized by a common parameter $t \in [0,1]$.
2. **Detection.** In detection tasks, the output is a (finite) list storing the positions of k corresponding landmarks. Thus, $GT = \{GT^i\}_{i=1}^k$, $AO = \{AO^i\}_{i=1}^k$, for $GT^i = (GT_j^i)_{j=1}^n$, $AO^i = (AO_j^i)_{j=1}^n$ points in $\mathbb{R}^n$, where n=1,2,3 is the dimension of the data domain.

Our final goal is to control (predict) the values taken by GT from the values taken by the alternative measure AO. In inference statistics, this is achieved by relating both quantities using a regression model.

2.1 Regression Model

The linear regression of a response variable y over an explicative variable x is given by:

$$Y = X\beta + \epsilon \tag{1}$$

for $\beta = (\beta_0, \beta_1)$ the regression parameters, $X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}$, $Y = (y_1, \cdots, y_N)$ a sampling of x, y and $\epsilon = (\varepsilon_1, \cdots, \varepsilon_N)$ an uncorrelated random error following a multivariate normal distribution, $N(0, \Sigma^2)$ of zero mean and variance $\Sigma^2 = \sigma^2 Id$.

The parameters of the regression model (1) are the regression coefficients $\beta = (\beta_0, \beta_1)$ and the error variance σ^2. The regression coefficients describe the way the two variables relate, while the variance indicates the accuracy of the model and, thus, measures to what extent x can predict y.

Given that, in our case, the inference is over GT, our model is:

$$GT_i = \beta_0 + \beta_1 AO_i + \varepsilon_i \tag{2}$$

for $(GT_i)_{i=1}^N$, $(AO_i)_{i=1}^N$ samplings of GT and AO obtained for each task as:

1. **Segmentation.** In the case of contours, the sampling is given by the coordinates of a uniform sampling of each of the curves:

$$\begin{array}{l} (GT_i)_{i=1}^{2N} = (GT(t_i))_{i=1}^N = (GT_1(t_i), GT_2(t_i))_{i=1}^N \\ (AO_i)_{i=1}^{2N} = (AO(t_i))_{i=1}^N = (AO_1(t_i), AO_2(t_i))_{i=1}^N \end{array}$$

   for $t_i = i/N, i = 1 : N$. In order to have pair-wise data, samplings are taken using a common origin of coordinates.
2. **Detection.** In this case, the sampling of the two variables is given by:

$$\begin{array}{l} (GT_i)_{i=1}^{N=nk} = (GT^i)_{i=1}^k = ((GT_j^i)_{j=1}^n)_{i=1}^k \\ (AO_i)_{i=1}^{N=nk} = (AO^i)_{i=1}^k = ((AO_j^i)_{j=1}^n)_{i=1}^k \end{array}$$

   Pair-wise data is obtained by using the same scanning direction in images for sorting the vector of landmarks.

For a sample of length N, the regression coefficients, $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1)$, are estimated by least squares fitting as:

$$\widehat{\beta} = (X^T X)^{-1} X^T Y \tag{3}$$

for X and Y as in eq. (1) and ^{T} denoting the transpose of a matrix.

The difference between the estimated response, $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$, and the observed response y_i, $e_i = y_i - \widehat{y}_i$, are called residuals. Their square sum provides an estimation of the error variance:

$$S_R = \widehat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$$

Previous to any kind of inference, it is mandatory to verify that the estimated parameters make sense. That is, whether it really exists a linear relation between x and y. By the Gauss-Markov theorem, such linear relation can be statistically checked using the following F-test [11]:

$$TM: \; H_0 : \beta_1 = 0 \; , \; H_1 : \beta_1 \neq 0 \tag{4}$$

where a $p - value$ close to zero (below α) ensures the validity of the linear model with a confidence $(1 - \alpha)100\%$.

2.2 Prediction Model

In order to predict the values of GT from the values achieved by AO, we use the regression prediction intervals [11]:

$$PI(x_0) = [L_{PI}(x_0), U_{PI}(x_0)]$$

since, for each $x = x_0$, they provide ranges for y at a given confidence level $1 - \alpha$. That is, given x_0, the values of the response y are within $L_{PI}(x_0) \leq y \leq U_{PI}(x_0)$ in $(1 - \alpha)100\%$ of the cases.

Given $x_0 = AO_0$, the confidence interval at a confidence level (1-α) predicting GT is given by:

$$PI(x_0) = [L_{PI}(x_0), U_{PI}(x_0)] = [\widehat{y_0} + t_{\alpha/2} S_R \sqrt{1 + h_0}, \widehat{y_0} - t_{\alpha/2} S_R \sqrt{1 + h_0}]$$

for $t_{\alpha/2}$ the value of a T-Student distribution with $N - 2$ degrees of freedom having a cumulative probability equal to $\alpha/2$ and $h_0 = (1 \quad x_0)(X^T X)^{-1}(1 \quad x_0)^T = a_0 + a_1 x_0 + a_2 x_0^2$. Prediction intervals achieve their minimum range at the average x and their maximum range at their extreme values xMin, xMax.

A prediction interval within a given precision, $U_{PI}(x) - L_{PI}(x) \leq \epsilon, \forall\, x$, indicates that the regression model predicts GT with high accuracy and, thus, AO is a good candidate for substituting GT. The alternative quantity can substitute GT in the measure that the identity line is within the range given by the prediction interval $PI(x)$. Otherwise, AO presents a systematic bias from the reference, which might be corrected using the regression coefficients. The slope, β_1, is associated to a scaling factor (unit change), while the intercept, β_0, is a constant bias.

The identity line is in the range of the prediction interval $PI(x)$ with a given precision, ϵ, if and only if $(PI(x) - x) \subset (-\epsilon, \epsilon)$, $\forall\ x$. This requirement is fulfilled if the following conditions hold:

$$\begin{array}{l} CP_1:\ max(L_{PI}(x) - x) \leq 0 \leq min(U_{PI}(x) - x) \\ CP_2:\ max(U_{PI}(x) - L_{PI}(x)) \leq 2\epsilon \end{array} \tag{5}$$

The first condition ensures that variables can be swapped with a confidence of $(1 - \alpha)$, while the second assesses the accuracy of the swapping. We note that the above conditions can also be formulated in terms of an identity test for the regression coefficients.

3 Results

We have chosen the following applications for each task:

1. **Vessel Wall Segmentation in Intravascular Ultrasound Sequences.** We have applied our model to the validation of the adventitia wall detection reported in [12] in order to compare the regression-prediction assessment to standard metrics (mean distance, noted by MeanD). We have considered two sequences of 300 frames each manually segmented every 20 frames (15 samples). One case (C1) has a low error and the other one (C2) a poor performance of the automatic method.
2. **Cardiac Phase Detection.** We have applied our model to assess replacing ECG signal sampling by manual sampling of longitudinal cuts of IntraVascular UltraSound sequences [13]. Comparison of cardiac phase samplings is a difficult task because it should not penalize constant shifts associated to a sampling of a different fraction of the cardiac phase. We have considered 3 sequences between 378 and 1675 frames long and acquisition rate between 10 and 30 fps. The first case (C1) is a short segment (378 frames) acquired without pullback. The other two are a (visually) good and bad acquisitions (C2 and C3, respectively).

Our goal is by no means validating the performance of alternative methods, but to show the benefits of regression-prediction models for performance evaluation. To such end, we have assessed the validity of the linear model (given by S_R and TM test), as well as, its prediction value (given by CP_i, $i = 1, 2$). Positions are given in mm.

Tables 1 and 2 report regression parameters and predictive value for each task and, in the case of segmentation (table 1), we also report the range (computed for 300 frames) of the metric MeanD in the last column. For the regression model, we report the $p - value$ for TM test, confidence intervals for β_0, β_1 and S_R. For the prediction model, we give the interval for the interchangeability condition, CP_1, and the accuracy ϵ in mm, CP_2. In the case of detection, CP_1 has been computed for $x + \beta_0$ instead of x in order to account for constant shift in samplings.

For all cases and tasks, there is a clear linear relation between GT and the image processing AO (with p close to the working precision). For the segmentation task (table 1), C1 has an accurate regression model close to the identity line. The squared root of the model accuracy ($\sqrt{S_R} = 0.1224$) agrees with MeanD ranges computed for the 15 manually segmented samples. Concerning predictive value, manual and automatic

Table 1. Regression-Prediction Model for Vessel Wall Segmentation

	Regression Model				**Prediction Model**		**Distance**
	TM	β_1	β_0	S_R	CP_1	CP_2	MeanD
C1	$\leq 10^{-308}$	1.009 ± 0.002	0.063 ± 0.001	0.015	(-0.072,0.280)	0.200	0.105 ± 0.018
C2	$\leq 10^{-308}$	1.001 ± 0.006	0.229 ± 0.029	0.221	(-0.561,0.899)	0.822	0.365 ± 0.171

Table 2. Regression-Prediction Model Scores for Cardiac Phase Detection

	Regression Model				**Prediction Model**	
	TM	β_1	β_0	S_R	CP_1	CP_2
C1	$\leq 10^{-308}$	0.998 ± 0.002	0.018 ± 0.038	0.004	(-0.113 ,0.047)	0.113
C2	$\leq 10^{-308}$	0.997 $\pm 7.5e^{-4}$	-0.284± 0.026	0.003	(-0.096,0.001)	0.094
C3	$\leq 10^{-308}$	0.966 ±0.004	0.792 ± 0.255	0.662	(-1.454,-2.449)	1.362

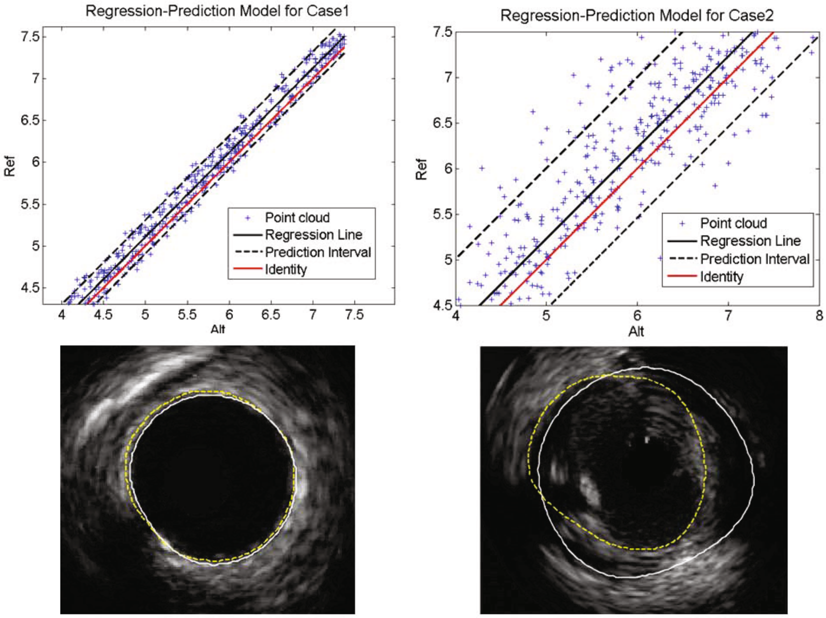


Fig. 1. Regression Model and Prediction Intervals for Vessel Wall Segmentation

contours can be swapped for the whole sequence with high accuracy. For C2, the model is a translation of the identity and the fitting is worse. This indicates severe contour misalignment (see right image in fig. 1). We observe that in this case the squared root of the fitting error ($\sqrt{S_R} = 0.4701$) also agrees with MeanD ranges. Although the two variables can be swapped (both measure the same [10]), the low accuracy of the

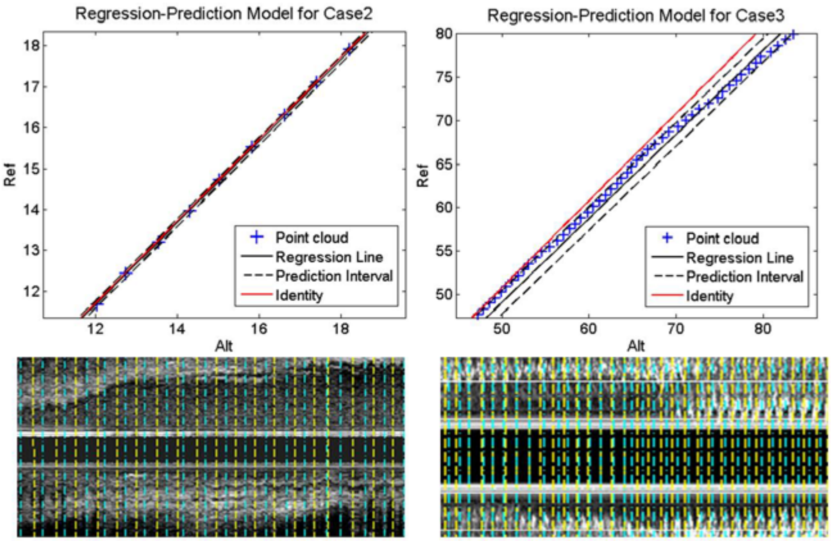


Fig. 2. Regression Model and Prediction Intervals for Cardiac Phase Detection

prediction advises against swapping them. For the detection task (table 2), C1 and C2 present an accurate regression model close to the identity line and a good predictive value. The non-zero intercept of C2 is due to a constant shift in manual samplings (see left image in fig. 2). Concerning C3, both, the regression model and the predictive value, present very bad scores and the samplings cannot be swapped (CP_1 does not hold).

Figures 1 and 2 show the regression-prediction plots for vessel wall segmentation and cardiac phase sampling, respectively. Each plot shows the point cloud AO (x-axis) versus GT (y-axis), the regression line in black solid, PI limits in dashed black and the identity line $AO = GT$ in red. In the case of vessel wall segmentation (fig. 1), we show a representative frame with manual (solid white) and automated (dashed yellow) contours, while for cardiac phase sampling (fig. 2) we show a longitudinal cut with ECG (yellow lines) and manual (cyan lines) samplings. For segmentation cases, the deviation of the identity line from the regression model is similar in both cases, though the range of the prediction interval is substantially larger for C2. This increase in error is reflected in the visual quality of the segmentation shown at the right bottom image. Regarding detection plots, visual inspection of the longitudinal sampling for C2 reasserts the agreement up to a constant shift reflected by the thin prediction interval in top left plots. For C3, the identity line traverses prediction interval upper bound as suggested by CP_1 interval. The prediction model coincides with the erratic relation between manual and ECG samplings observed in the left bottom image.

4 Conclusions and Future Work

Standardized validation of medical imaging algorithms allowing generalization of conclusions to clinical data is a challenging task not fully solved. We have approached validation from the point of view of statistical inference. In this context, we use a regression model for assessing to what extent GT and AO can be swapped and a prediction model for inferring conclusions to the whole population. Experiments on a segmentation task are a good proof of concept of the capability of the framework for assessing performance, while experiments on a detection task illustrate its versatility.

The framework presented in this paper can be applied to explore the performance from a relatively small test set. In order to fully generalize results to the whole clinical data involved in each task, we should consider a general regression model with random effects in order to account for variability across acquisitions. Also, in order to fully validate their capability for assessing performance, we are running our methods on the whole data set and metrics used in [12].

Acknowledgements. We thank InCor from USP, São Paulo, for their support with the IVUS images database. This work was supported by the Spanish projects PI071188, TIN2009-13618, TRA2010-21371-C03-01, CSD2007-00018 and the 2nd author by The Ramon y Cajal Program.

References

1. Jannin, P., Krupinski, E., Warfield, S.: Guest editorial validation in medical imaging processing. IEEE Trans. on Med. Imag. 25(11), 1405–1409 (2006)
2. Wiest-Daesslé, N., Prima, S., Morrissey, S.P., Barillot, C.: Validation of a new optimisation algorithm for registration tasks in medical imaging. In: ISBI 2007, pp. 41–44 (2007)
3. Lee, S., Abràmoff, M.D., Reinhardt, J.M.: Validation of retinal image registration algorithms by a projective imaging distortion model. In: Conf. Proc. IEEE Eng. Med. Biol. Soc. 2007, pp. 6472–6475 (2007)
4. Jannin, P., Fitzpatrick, J., Hawkes, D., Pennec, X., Shahidi, R., Vannier, M.: Validation of medical image processing in image-guided therapy. IEEE Trans. Med. Imag. 21(12), 1445–1449 (2002)
5. Gee, J.: Performance evaluation of medical image processing algorithms. In: Proc. SPIE, vol. 3979, pp. 19–27 (2000)
6. Castro, F., Pollo, C., Meuli, R., Maeder, P., Cuisenaire, O., Cuadra, M., Villemure, J.G., Thiran, J.P.: A cross validation study of deep brain stimulation targeting: From experts to atlas-based, segmentation-based and automatic registration algorithms. IEEE Trans. Med. Imag. 25(11), 1440–1450 (2006)
7. Landis, J., Koch, G.: The measurement of observer agreement for categorical data. Biometrics 33, 159–174 (1977)
8. Gerig, G., Jomier, M., Chakos, M.: Valmet: A new validation tool for assessing and improving 3D object segmentation. In: Niessen, W.J., Viergever, M.A. (eds.) MICCAI 2001. LNCS, vol. 2208, pp. 516–528. Springer, Heidelberg (2001)
9. Bland, J.M., Altman, D.G.: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1(8476), 307–310 (1986)
10. Hoppin, J., Kupinski, M., Kastis, G., Clarkson, E., Barrett, H.: Objective comparison of quantitative imaging modalities without the use of a gold standard. IEEE Trans. Med. Imag. 21(5), 441–449 (2002)

11. Newbold, P., Carlson, W.L., Thorne, B.: Statistics for Business and Economics, 6th edn. Pearson Education, London (2007)
12. Gil, D., Hernàndez, A., Rodriguez, O., Mauri, J., Radeva, P.: Statistical strategy for anisotropic adventitia modelling in IVUS. IEEE Trans. Med. Imag. 25(6), 768–778 (2006)
13. Hernàndez-Sabaté, A., Gil, D., Garcia-Barnés, J., Martí, E.: Image-based cardiac phase retrieval in Intravascular Ultrasound sequences. IEEE Trans. Ultr., Ferr., Freq. Ctr. 58(1), 60–72 (2011)

Glaucoma Classification Based on Histogram Analysis of Diffusion Tensor Imaging Measures in the Optic Radiation

Ahmed El-Rafei[1,4], Tobias Engelhorn[2], Simone Wärntges[3], Arnd Dörfler[2], Joachim Hornegger[1,4], and Georg Michelson[3,4,5]

[1] Pattern Recognition Lab, Department of Computer Science
ahmed.el-rafei@informatik.uni-erlangen.de,
joachim.hornegger@informatik.uni-erlangen.de
[2] Department of Neuroradiology
tobias.engelhorn@uk-erlangen.de,
arnd.doerfler@uk-erlangen.de
[3] Department of Ophthalmology
simone.waerntges@uk-erlangen.de,
georg.michelson@uk-erlangen.de
[4] Erlangen Graduate School in Advanced Optical Technologies (SAOT)
[5] Interdisciplinary Center of Ophthalmic Preventive Medicine and Imaging (IZPI),
Friedrich-Alexander University Erlangen-Nuremberg, Germany

Abstract. Glaucoma is associated with axonal degeneration of the optic nerve leading to visual impairment. This impairment can progress to a complete vision loss. The transsynaptic disease spread in glaucoma extends the degeneration process to different parts of the visual pathway. Most of glaucoma diagnosis focuses on the eye analysis, especially in the retina. In this work, we propose a system to classify glaucoma based on visual pathway analysis. The system utilizes diffusion tensor imaging to identify the optic radiation. Diffusion tensor-derived indices describing the underlying fiber structure as well as the main diffusion direction are used to characterize the optic radiation. Features are extracted from the histograms of these parameters in regions of interest defined on the optic radiation. A support vector machine classifier is used to rank the extracted features according to their discrimination ability between glaucoma patients and healthy subjects. The seven highest ranked features are used as inputs to a logistic regression classifier. The system is applied to two age-matched groups of 39 glaucoma subjects and 27 normal controls. The evaluation is performed using a 10-fold cross validation scheme. A classification accuracy of 81.8% is achieved with an area under the ROC curve of 0.85. The performance of the system is competitive to retina based classification systems. However, this work presents a new direction in detecting glaucoma using visual pathway analysis. This analysis is complementary to eye examinations and can result in improvements in glaucoma diagnosis, detection, and treatment.

Keywords: Classification, Diffusion Tensor Imaging, Optic Radiation, Glaucoma, Histogram, Visual System.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 529–536, 2011.

1 Introduction

More than 60 million people around the world suffer from glaucoma. Bilateral blindness caused by glaucoma is estimated to affect more than 8 million people [1]. Glaucoma is accompanied by neurodegeneration of the axonal fibers in the optic nerve along with visual impairment. The development of glaucoma can result in complete blindness. The vision loss can not be restored. However, if glaucoma is detected in an early stage, its progression can be delayed or stopped. Therefore, early detection of glaucoma is necessary as well as novel treatment methods.

The conventional trend in glaucoma diagnosis is through eye examinations. Intraocular blood pressure, retinal nerve fiber layer thickness measured by optical coherence tomography (OCT), fundus images, and optic disc topography evaluated by Heidelberg retina tomograph (HRT) are examples of glaucoma relevant data examined by ophthalmologists to evaluate the glaucoma severity. Moreover, systems were developed based on the aforementioned data among others using various eye imaging modalities to screen, detect, and diagnose glaucoma [2,3]. Despite the efficiency and high performance of the developed systems, they focus on the eye, specifically the retina, ignoring the largest part of the visual system represented by the cerebral visual pathway fibers within the brain. In addition, the mechanism of glaucoma progression and the functional or structural damage precedence [4] are still unresolved issues. Therefore, exploring the recently discovered possibilities offered by diffusion tensor imaging (DTI) [5] to reconstruct and characterize the fiber structure of the human white matter [6] can be a valuable addition to the glaucoma examination flow.

Recent studies addressed the visual system changes due to glaucoma. Garaci et al. [7] showed that a reduction in fiber integrity affecting different parts of the visual pathway as the optic nerve and optic radiation is correlated with glaucoma. Another study showed axonal loss along the visual pathway from the optic nerve through the lateral geniculate nucleus till the visual cortex in the presence of glaucoma [8]. These results suggest that the visual pathway analysis can be significant in detecting and diagnosing glaucoma.

In this article, we investigate the significance of DTI-derived parameters in the optic radiation for glaucoma detection. We propose a classification system based on statistical features derived from the histograms of the DTI indices. The optic radiation is first identified automatically using the authors' developed algorithm [9]. A specific region of interest (ROI) on the optic radiation is then manually delineated. The histograms of the DTI measures are calculated. The histograms' statistical features are extracted from the histograms of the DTI indices in the specified ROI. The features are evaluated using a support vector machine classifier for dimensionality reduction and the highest ranked features are used for classification. The system is trained and tested using 10-fold cross validation. Finally, the ability of the system to differentiate between normal subjects and glaucoma patients is evaluated.

2 Classification System

2.1 Diffusion Tensor Imaging

Diffusion-weighted imaging (DWI) brain scans were acquired using a 3T-MRI high field scanner (Magnetom Tim Trio, Siemens, Erlangen, Germany). The diffusion weighting gradients were applied along 20 non-collinear directions with a maximal b-factor of 1,000 s/mm^2. The scans were repeated four times and averaged to increase the signal to noise ratio (SNR) and to improve the quality of the images. The axial resolution was 1.8 × 1.8 mm^2 with 5 mm slice thickness. The corresponding acquisition matrix size was 128 × 128 on a field of view (FoV) of 23 × 23 cm^2. The acquisition sequence protocol was a single-shot, spin echo, echo planar imaging (EPI) with parameters: TR = 3400 ms, TE = 93 ms, and partial Fourier acquisition = 60%. The scans were complemented by a non-weighted diffusion scan with b-factor equals zero. The Gaussian modeling of the diffusion process within a voxel is represented by a 3 × 3 diffusion tensor. The diffusion tensors were calculated from the DWI-datasets. The eigenvalue decomposition of the diffusion tensors contained information about the principal diffusion direction and aspects of the diffusion process (degree of anisotropy, mean diffusion, etc). The diffusion tensors were spectrally decomposed. The obtained eigenvalues were used to calculate the mean (MD), radial (RD), and axial (AD) diffusivities in addition to the fractional anisotropy (FA) [10]. The eigenvector corresponding to the largest eigenvalue was regarded as the principal diffusion direction (PDD).

2.2 Optic Radiation Segmentation

The identification of the optic radiation was performed using the authors' previously developed algorithm [9]. The algorithm operated on the interpolated DTI-images to produce an automatic segmentation of the optic radiation. The drawbacks of the Euclidean space interpolation and analysis of diffusion tensors were avoided by the utilization of the Log-Euclidean framework [11]. The DTI-images were enhanced by applying an anisotropic diffusion filtering to the individual elements of the diffusion tensors. This increased the coherency within the fiber bundles while preserving their edges. Based on neurophysiological facts of the dominant diffusion direction in the optic radiation and its anatomical size relative to other fibers, the optic radiation was initially identified using a thresholding and connectivity analysis. Similarly, the mid brain was approximately identified to be used later for segmentation enhancement. A region-based segmentation with the initialization of the optic radiation from the previous step was performed by a statistical level set engine [12]. The level set segmentation was adjusted to work with the Log-Euclidean metric for extending the framework to Riemannian operations while maintaining the computational efficiency. The framework optimized the posterior probabilities of partitioning the brain image space into the optic radiation and the remaining parts of the brain. The probabilities were modeled by normal distributions of the diffusion tensors within each of the two division parts. Finally, the outcome of the level set segmentation was adjusted based on the relative anatomical position between the optic radiation and the mid brain. This was done to remove the tracts anteriorly connected to the optic radiation (i.e., optic tracts). Further details on the segmentation system can be found in [9].

2.3 Region of Interest Selection

In this step, a region of interest defined on the segmented optic radiation was configured. The slice containing the optic radiation and clearly identifying the termination of the optic tracts in the lateral geniculate nucleus (LGN) region was located in all subjects. The automatic segmentation on the selected slice was examined by two DTI experts and the segmentation errors were manually corrected. Moreover, the connection of the optic radiation to the primary visual cortex was manually eliminated. This region is characterized by misleading reduced fractional anisotropy due to the limitation of the diffusion tensor in modeling the branching and crossing fibers [13]. The final processed optic radiation on the selected slice was the ROI used in the remaining analysis. Figure 1 shows an example of a selected ROI on a sample subject.

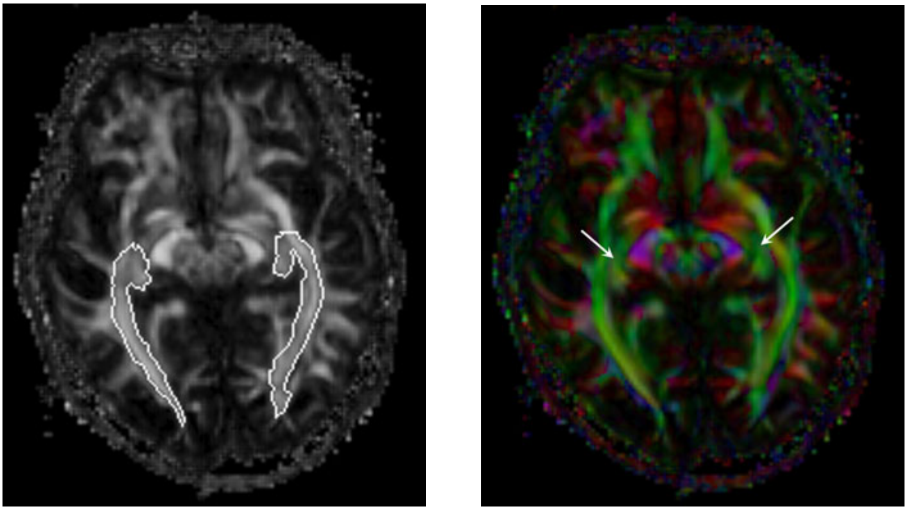


Fig. 1. The semi-automatically identified region of interest (ROI) representing the optic radiation shown on a fractional anisotropy image (left). The diffusion direction coded image (right) of the ROI-slice demonstrates the dominant anterior-posterior diffusion direction in the optic radiation. The selected slice indicates clearly the termination of the optic tracts at the lateral geniculate nuclei (LGN) as indicated by the white arrows on the right image.

2.4 Histogram Analysis and Feature Extraction

The histograms of the four DTI-derived parameters (FA, MD, RD, and AD) in the specified ROI were computed. A number of bins for each parameter were predetermined and the number of voxels corresponding to a certain bin range was calculated. The PDD has a unity length with three components representing the three coordinate axes. The PDD was converted to the spherical coordinate system. The histograms of the azimuth and inclination angles were measured by binning them in 0.2 radians bins. Since the sign of the PDD is not representative, the range of the azimuth angle was restricted between zero and 180 degrees while the inclination angle range was retained between zero and 180 degrees. That was simply done by inverting the direction of the PDD if it falls outside these ranges. Six first order statistical features (Mean, variance, skewness, kurtosis, energy, and entropy) of the DTI-indices and the PDD were derived from the histograms using the following equations:

$$Mean : \mu = \sum_{i=1}^{N} param(i) \times hist(i) \tag{1}$$

$$Variance : \sigma^2 = \sum_{i=1}^{N} (param(i) - \mu)^2 \times hist(i) \tag{2}$$

$$Skewness : \mu_3 = \sigma^{-3} \sum_{i=1}^{N} (param(i) - \mu)^3 \times hist(i) \tag{3}$$

$$Kurtosis : \mu_4 = \sigma^{-4} \sum_{i=1}^{N} (param(i) - \mu)^4 \times hist(i) - 3 \tag{4}$$

$$Energy : E = \sum_{i=1}^{N} [hist(i)]^2 \tag{5}$$

$$Entropy : H = -\sum_{i=1}^{N} hist(i) \log(hist(i)) \tag{6}$$

where N is the number of bins in the corresponding DTI-parameter histogram, *hist* is the normalized histogram (i.e. probability distribution which is the histogram divided by the total number of voxels within the ROI), i is the index of the ith bin, and $param(i)$ is the mean value of the corresponding parameter (param) in the ith bin.

2.5 Feature Selection and Classification

A support vector machine classifier [14] was used to rank the 36 histogram features by recursive feature elimination. This procedure works as follows: The support vector machine classifier was trained using the complete feature set and the features' weights were determined. Then, the feature with the lowest squared weight was considered as the least ranked feature. The feature with the lowest rank was removed from the feature set. The previous steps were repeated iteratively with the remaining features until all the features were ranked. The highest seven ranked features provided the best classification performance and were, therefore, selected as features for the classifier. For classification, the selected seven features were the input to a logistic regression classifier. The training and testing were performed using a 10-fold cross validation analysis. The software implementation in Weka [15] was used for the feature selection and the classification.

3 Results

The proposed system was applied to two groups of subjects: A group of 27 healthy controls with a mean age of 58.52 ± 10.10 years (17 females and 10 males) and 39 patients with primary open angle glaucoma (POAG) with a mean age of 61.74 ± 8.32 years (19 females and 20 males). The two groups were age matched and the two-sided Wilcoxon ranksum test which is equivalent to the Mann-Whiteny U-test gave a p-value of 0.17 indicating the correlation between the ages of the two groups. The subjects underwent MRI and DTI brain scans. The brains were examined by experienced neuroradiologists and did not show any indications of neuronal diseases or lesions affecting the visual pathway. The optic radiations of

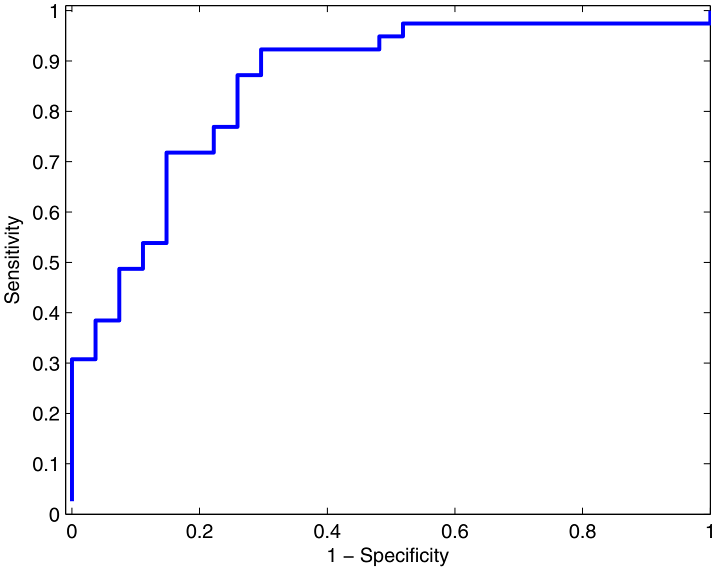


Fig. 2. The Receiver Operating Characteristic (ROC) curve of the glaucoma classification system based on DTI measures. The area under the ROC curve is 0.85.

all subjects were segmented and the ROIs were selected. The statistical features were extracted from the histograms of the four DTI-derived indices as well as the azimuth and inclination angles of the PDD. The features were ranked by a support vector machine classifier. The seven most discriminating features were: MD Kurtosis, RD Skewness, FA Entropy, MD Skewness, Azimuth Energy, Azimuth Entropy, and FA Mean, respectively. A logistic regression classifier was trained and tested using these seven features in a 10-fold cross validation setup.

The classification accuracy of the system was 81.82%. This rate corresponded to the correct recognition of 54 subjects' classes. Out of these 54 subjects, 36 were glaucoma patients and 18 were control subjects. Three glaucoma patients and 9 normal subjects were wrongly diagnosed. The receiver operating characteristic (ROC) curve was calculated and plotted in Figure 2. The area under the ROC curve was 0.853. A sensitivity of 92.31% for glaucoma detection and specificity of 70.37% were obtained. Additional values from the ROC curve at a different threshold showed a sensitivity of 71.79% at a fixed specificity of 85.19%.

4 Discussion and Conclusion

This paper proposed a new approach in glaucoma detection using visual pathway analysis. Utilizing the capabilities of the diffusion tensor imaging, the system identified and characterized the fiber structure of the optic radiation. First order statistical features extracted from the histograms of the DTI-derived measures were used to detect glaucoma. The classification performance obtained by the proposed system is comparable to systems based on eye imaging modalities [3]. Nevertheless, the significance of the DTI-parameters and the histogram features is evident from the limited number of features used.

Diffusion tensor-derived indices characterize different aspects of the underlying fiber structure. For example, FA indicates the degree of intravoxel fiber alignment and coherency while MD is related to the fiber integrity. Thus, these parameters were shown to correlate with the cerebral fiber damage caused by neuronal diseases such as Alzheimer and glaucoma. Four classification features among the highest ranked features were derived from the FA and the MD histograms demonstrating the sensitivity of these parameters to glaucoma. Fractional anisotropy and MD were shown to correlate with glaucoma [7] and such an influence can be expected.

The proposed classification method based on visual pathway analysis presents a new perspective in detecting diseases affecting the visual system such as glaucoma. Diffusion tensor imaging provides valuable information regarding the white matter microstructure allowing for the identification, characterization, and pathological diagnosis of fiber tracts. The high classification rates are indicators of the sensitivity of the features derived from the DTI-measures to glaucoma. It also emphasizes the effect of glaucoma on the entire visual system. This analysis is complementary to retina-based diagnosis. The integration of features from traditional eye imaging modalities and diffusion tensor imaging covers the complete visual system. Thus, it can enhance the detection of glaucoma significantly, the understanding of its pathophysiology, and consequently the treatment methods.

Acknowledgements. The authors gratefully acknowledge funding of the Erlangen Graduate School in Advanced Optical Technologies (SAOT) by the German National Science Foundation (DFG) in the framework of the excellence initiative.

References

1. Quigley, H.A., Broman, A.T.: The number of people with glaucoma worldwide in 2010 and 2020. The British Journal of Ophthalmology 890(3), 262–267 (2006)
2. Burgansky-Eliash, Z., Wollstein, G., Bilonick, R.A., Ishikawa, H., Kagemann, L., Schuman, J.S.: Glaucoma detection with the Heidelberg retina tomograph 3. Ophthalmology 114(3), 466–471 (2007)
3. Bock, R., Meier, J., Nyúl, L.G., Hornegger, J., Michelson, G.: Glaucoma risk index: Automated glaucoma detection from color fundus images. Medical Image Analysis 14(3), 471–481 (2010)
4. Hood, D.C., Kardon, R.H.: A framework for comparing structural and functional measures of glaucomatous damage. Prog. Retin. Eye. Res. 26(6), 688–710 (2007)
5. Basser, P.J., Mattiello, J., Lebihan, D.: MR diffusion tensor spectroscopy and imaging. Biophysical Journal 66(1), 259–267 (1994)
6. Staempfli, P., Rienmueller, A., Reischauer, C., Valavanis, A., Boesiger, P., Kollias, S.: Reconstruction of the human visual system based on DTI fiber tracking. Journal of Magnetic Resonance Imaging 26(4), 886–893 (2007)
7. Garaci, F.G., Bolacchi, F., Cerulli, A., Melis, M., Spanó, A., Cedrone, C., Floris, R., Simonetti, G., Nucci, C.: Optic nerve and optic radiation neurodegeneration in patients with glaucoma: in vivo analysis with 3-T diffusion-tensor MR imaging. Radiology 252(2), 496–501 (2009)
8. Gupta, N., Ang, L.C., de Tilly, L.N., Bidaisee, L., Yücel, Y.H.: Human glaucoma and neural degeneration in intracranial optic nerve, lateral geniculate nucleus, and visual cortex. British Journal of Ophthalmology 90(6), 674–678 (2006)
9. El-Rafei, A., Engelhorn, T., Waerntges, S., Doerfler, A., Hornegger, J., Michelson, G.: Automatic segmentation of the optic radiation using DTI in healthy subjects and patients with glaucoma. In: Tavares, J.M.R.S., Jorge, R.M.N. (eds.) Computational Vision and Medical Image Processing. Computational Methods in Applied Sciences, vol. 19, pp. 1–15. Springer, Netherlands (2011)
10. Basser, P.J., Pierpaoli, C.: Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. Journal of Magnetic Resonance Series B 111(3), 209–219 (1996)
11. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Log-Euclidean metrics for fast and simple calculus on diffusion tensors. Magn. Reson. Med. 56(2), 411–421 (2006)
12. Lenglet, C., Rousson, M., Deriche, R.: DTI segmentation by statistical surface evolution. IEEE Transactions on Medical Imaging 25(6), 685–700 (2006)
13. Tuch, D.S., Reese, T.G., Wiegell, M.R., Makris, N., Belliveau, J.W., Wedeen, V.J.: High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity. Magnetic Resonance in Medicine 48(4), 577–582 (2002)
14. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning 46(1-3), 389–422 (2002)
15. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1), 10–18 (2009)

Textural Classification of Abdominal Aortic Aneurysm after Endovascular Repair: Preliminary Results

Guillermo García[1], Josu Maiora[2], Arantxa Tapia[1], and Mariano De Blas[3]

[1] Engineering Systems and Automatic Department –EUP, University of the Basque Country, San Sebastián, Spain

[2] Electronics and Telecomunications Department –EUP, University of the Basque Country, San Sebastián, Spain

[3] Interventional Radiology Department, Donostia Hospital, San Sebastián, Spain

Abstract. In recent years, endovascular aneurysm repair (EVAR) has proved to be an effective technique for the treatment of abdominal aneurysm. However, complications as leaks inside the aneurysm sac (endoleaks) can appear, causing pressure elevation and increasing the danger of rupture consequently. Computed tomographic angiography (CTA) is the most commonly used examination for medical surveillance, but endoleaks can not always be detected by visual inspection on CTA scans. The aim of this work was to evaluate the capability of texture features obtained from CT images, to discriminate evolutions after EVAR. Regions of interest (ROIs) from patients with different post-EVAR evolution were extracted by experienced radiologists. Three different techniques were applied to each ROI to obtain texture parameters, namely the gray level co-occurrence matrix (GLCM) , the gray level run length matrix (GLRLM) and the gray level difference method (GLDM). In order to evaluate the discrimination ability of textures features, each set of features was applied as input to support vector machine (SVM) classifier. The performance of the classifier was evaluated using 10-fold cross validation with the entire dataset. The average of accuracy, sensitivity, specificity, receiving operating curves (ROC) and area under the ROC curves (*AUC*) were calculated for the classification performances of each texture-analysis method. The study showed that the textural features could help radiologists in the classification of abdominal aneurysm evolution after EVAR.

Keywords: Aneurysm, EVAR, texture features, support vector machine.

1 Introduction

The *Endovascular Aneurysm Repair* (EVAR) treatment, is a percutaneous image-guided endovascular procedure in which a stent graft is inserted into the aneurysm cavity. Once the stent is placed, the blood clots around the metallic mesh forcing the blood flux through the stent and thus reducing the pressure on the aneurysm walls. Nevertheless, in a long term perspective different complications such as prostheses displacement or leaks inside the aneurysm sac (endoleaks) could appear provoking a pressure elevation and increasing the danger of rupture consequently. Due to this,

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854 pp. 537–544, 2011.

periodic follow-up scans of the prosthesis behaviour are necessary. At present, contrast enhanced computed tomographic angiography (CTA) is the most commonly used examination for imaging surveillance [1]. On the other hand, the post operation analysis is quite crude as it involves manually measuring different physical parameters of the aneurysm cavity [2]. According to these measurements, the evolution of the aneurysm can be split up into two main categories: *Favourable evolution*, when a reduction of the diameter of the aneurysm sac can be observed what means that the aneurysm has been correctly excluded from the circulation. *Unfavourable evolution*, when a growth of the aneurysm diameter in presence of endoleaks is observed. Endoleaks can be detected thanks to contrast in the CT images.

In Fig. 1 two series of images for favourable and unfavourable evolution are shown. We could also distinguish a subcategory inside the unfavourable evolution cases. There are patients in which abdominal aneurysm does not increase or reduce significantly its volume and endoleaks are not visually detected (endotensión) [3]. The reason for this behaviour it is not completely known but it is usually attributed to different causes [4]. The initial idea behind the study is that texture thrombus in favourable shrinking aneurysms might differ from unfavourable expanding ones. If the hypothesis is confirmed, we consider to extend the analysis to endotension cases in posterior studies.

1.1 Texture Analysis

In recent years, many efforts have been put into the developing of Computer Aided Diagnosis systems based on image processing methods. The principal motivation for the research on this kind of systems has been to assist the clinicians on the analysis of medical images. In many occasions this analysis implies the detection or measurement of subtle differences, usually difficult to appreciate by visual inspection even for experienced radiologists. Computer Aided Diagnosis systems have been successfully utilized in a wide range of medical applications [5-7]. A particular field inside the image processing methods is the so named texture-based analysis. This analysis studies, not only the variation of the pixel intensity values along the image but also the possible spatial arrangement of them, and the more or less periodic repetition of such arrangement (primitives). From this point of view texture analysis can help on the functional characterization of different kind of organs, tissues, etc, at the evolution of disease. The textures features obtained from the analysis can be fed as inputs for a deterministic or probabilistic classifier, which assign each sample with its specific class.

Textures analysis methods can generally be classified into three categories: statistical methods, model based methods, and structural methods [8]. In our approach we have focused on the application of statistical texture methods, specifically, on spatial domain statistical techniques as the Gray Level Co-occurrence Matrix (GLCM) [9], the Gray Level Run Length Matrix (GLRLM) [10] and the Gray Level Difference Method (GLDM) [11]. These three very extended methods can capture second or higher order statistics on the relation between gray values in pixel pairs or groups of pixels in order to estimate their probability-density functions. Their validity has been proved in many studies [12-14].

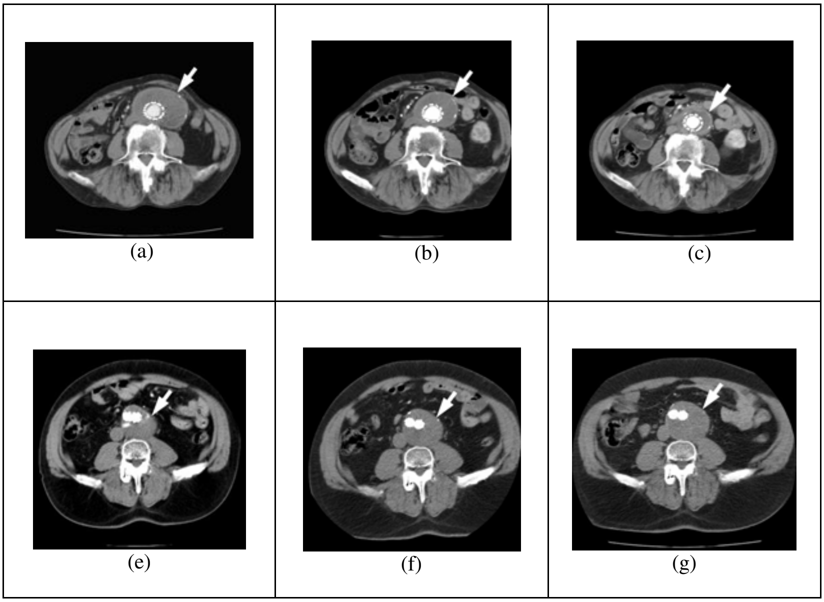


Fig. 1. CTA images of 2 patients treated with EVAR. Top row: favourable evolution. (a) 1 year after treatment. (b) 2 years after. (c) 3 years after. Bottom row: unfavourable evolution.(e) 1 year after treatment. (f) 2 years after. (g) 3 years after. White arrow points aneurysm sac in all of the scans.

Our purpose in this study is to investigate the GLCM, GLRLM and GLDM capacity for discriminate between favourable and unfavourable evolutions of patients after abdominal EVAR treatment. For obtaining this objective a semi-automated segmentation process to facilitate the extraction of samples has been developed. Once the samples from patients with different post-EVAR evolution have been obtained, the texture features from the three methods are calculated and fed into a support vector machine classifier for automatic classification.

The paper is organised as follows: "Materials and Methods" provides with information about the acquisition of CTA images, the description of the segmentation process, the theoretical background on the texture analysis methods, and the definition of the support vector machine structure. The methodology, and results obtained from the performance evaluation of the classifier are presented in "Results". Finally, some conclusions are given in "Conclusion" section.

2 Materials and Methods

2.1 Dataset

The CTA image scans used in this work were obtained by experienced radiologists from the Vascular Surgery Unit and Interventional Radiology Department of the Donostia Hospital. These CTA images belong to the scan studies of 70 patients with ages ranging from 70 to 93 years, conducted over a maximum of 5 years in fixed periods of 6 -12 months. The total patient set was selected by the radiologist team and balanced samples sets for training and testing the classifier system were obtained. A total of 35 CTA studies belonged to the "favourable evolution" class and 35 to the "unfavourable" one. All the studies were taken from the abdominal area with a spatial resolution of 512×512 pixels and 12-bit gray-level at the WW400 y WL40 window and 5 mm thickness in DICOM format. For each patient a maximum of three ROIs (15x15 pixels) were extracted from different slices, resulting in a total of 210 ROIs. Half of them corresponded to the "favourable evolution" group and the rest to the "unfavourable" one.

2.2 Segmentation Process

In order to facilitate the extraction of samples by radiologists a semi-automated segmenting process of the aneurysm has been implemented [15]. Based on the series of CTA scans in DICOM, images are created with a volume format. During this process the resolution and the spacing of the original images are preserved. The files obtained are used as inputs for the following 3D processing pipeline. The spinal canal is segmented to use it as reference, as it deforms only a little and it is relatively easy to segment. The aneurysm is segmented using a combination of fast marching method to delineate the thrombus and confidence connected components to delineate the stent graft. The two segmentations are then fused together using geodesic active contours and smoothed using a median filtering. The segmented volume is resliced along the axial plane facilitating the extraction of the 15x15 pixels ROIs of thrombus aneurysm samples by specialists.

2.3 Texture Analysis – Feature Extraction

2.3.1 Gray Level Co-occurrence Matrix (GLCM)

The gray level co-occurrence matrix (GLCM) [9] is an estimation of a second order joint conditional probability density function $f(i,j\,/d,\ \theta)$. This function characterizes the spatial interrelationships of the gray values in an image. The values of the co-occurrence matrix elements represent the probability of going from grey level i to grey level j given that they are separated by the distance d and the direction is given by the angle θ (usually θ = 0°, 45°, 90°, and 135°). In the present application, GLCM features have been calculated at distance 1 due to the reduced size of the aneurysm samples. Initially, the assumption of an isotropic texture distribution inside the aneurysm sac was considered, consequently averaging over the four angular directions was computed. To reduce the influence of random noise on texture features, the number of gray levels was reduced to 16 prior to the accumulation of the matrix. From GLCM matrix a set of features are obtained to classify the kind of texture analysed. In this study, 13 features have been evaluated: Energy, correlation, inertia, entropy, inverse difference moment, sum average, sum variance, sum entropy, difference average, difference variance, difference entropy and two information measures of correlation.

2.3.2 The Gray Level Run Length Matrix (GLRLM)

The gray level run length matrix (GLRLM) [10] is a way of extracting higher order statistical texture information. For a given image, a gray level run is a set of linearly adjacent picture points having the same gray level value. A run length matrix is a two-dimensional matrix in which each element $p(i,j / \theta)$ represents the total number of runs with pixels of gray value i and run length j in a certain direction θ. The number of gray levels in the image is usually reduced by re-quantization before the accumulation of the matrix. In our study the number of gray levels has been kept in 16, equal than in the GLCM method in order to make both methods comparable. Various textures features can then be extracted from the run length matrix. In our case the following 11 features has been calculated: short run emphasis, long run emphasis, gray-level nonuniformity, run length nonuniformity, run percentage, low gray-level run emphasis, high gray-level run emphasis, short run low gray-level emphasis, short run high gray-level emphasis, long run low gray-level emphasis, and long run high gray-level emphasis.

2.3.3 Gray Level Difference Method

The Gray Level Difference method (GLDM) [11] is based on the occurrence of two pixels which have a given absolute difference in gray level and which are separated by a specific displacement δ, estimated by the probability-density function D(i /δ). In this analysis, four possible forms of the vector δ will be considered: (0, d), (-d , d), (d, 0), and (d , -d) where d is the intersample spacing distance. Due to the reduced size of the aneurysm samples, d distance was considered equal to 1. Five textural features are measured from D(i /δ): contrast, angular second moment, entropy, mean, and inverse difference moment. As with the GLCM method, the assumption of an isotropic texture distribution inside the aneurysm sac was considered, consequently averaging over the four angular directions was computed.

2.4 Classification

To test the approach, a Support Vector Machine (SVM) [16] has been utilized as texture classifier. Support vector machine classifier is a powerful machine learning tool which that generalizes well to a wide range of real-world applications [17-18]. The basic idea of SVM classifier is to determine a separating hyperplane that distinguishes between two classes. Given a set of labeled training data, the input data are transformed into high-dimensional feature space with the use of kernel functions, so that the transformed data becomes more separable compared to the original ones. The SVM attempts to minimize a bound on the generalization error and therefore it tends to perform well when applied to data outside the training set. Lots of variety kernels can be employed for mapping input data but some of the most frequently used kernel functions are Gaussian, Polynomial and Sigmoid kernels. In this study, the Gaussian radial basis function (RBF) kernel was utilized because of its good results in many practical applications. Three SVMs with different number of inputs were implemented. A SVM with 13 inputs for GLCM, another with 11 inputs for GLRLM, and another one with 5 inputs for GLDM. All the textural features were normalized by the sample mean and standard deviation of the data set before being fed to the

SVM. In our case, the best classification was obtained with a penalty factor value of +4.7 and a sigma value of +12.5 for cross correlation protocol.

3 Results

In order to evaluate the potential of texture analysis to discriminate between the two types of aneurysm evolution the development and validation of the SVM has been based on the 10-fold cross validation method. To improve the evaluation of performance of the feature sets, the process was repeated 6 times averaging the results.The averages of accuracy, sensitivity, and specificity for all the trials were used as an estimate of the performance of each classifier and consequently of the discrimination ability of textures features for differentiate between favourable or unfavourable cases. Table 1 shows the accuracy, sensitivity, and specificity values (mean ± standard deviation) estimated by the cross validation of the testing sets for each texture method.

Table 1. The average classification accuracy, sensitivity and specificity (in %) for testing sets of the SVMs are given for the GLCM, GLDM, and GLRLM features

Texture Method	Accuracy (%)		Sensitivity (%)		Specificity (%)	
	mean	std	mean	std	mean	std
GLCM	93.61	0.15	95.08	0.12	90.80	0.24
GLDM	91.41	0.23	93.20	0.21	90.20	0.17
GLRLM	84.05	0.47	85.44	0.17	81.74	0.19

From the table 1, it is shown that all texture analysis methods supplied the support vector machine classifiers with enough discriminative information to differentiate between aneurysm evolutions. The best performing features set in terms of correctly classified cases corresponds to the GLCM method (93.64% ± 0.15) but the other two methods, GLDM (91.41% ± 0.23) and GLRLM (81.74% ± 0.19), could also be considered as significant. The area under the ROC curve (AUC) was also used as a measure of the classification performance. Table 2 presents the AUC values (mean ± standard deviation) calculated for feature training and testing sets using the 10-fold cross validation.

Table 2. The AUC values (mean ± standard deviation) calculated for feature testing sets using the 10-fold cross validation

Texture Method	AUC_mean testing	
	mean	std
GLCM	0.930	0.020
GLDM	0.883	0.032
GLRLM	0.821	0.042

The obtained values show again that the biggest area under the ROC curve, and consequently the best performance of the classifier is obtained with the set of features extracted with the GLCM method (0.930 ± 0.020), followed by the GLDM (0.883± 0.032) and the GLRLM methods (0.821± 0.042). Figure 4 depicts the average ROC curves obtained using the 10-fold cross validation testing sets for each texture method.

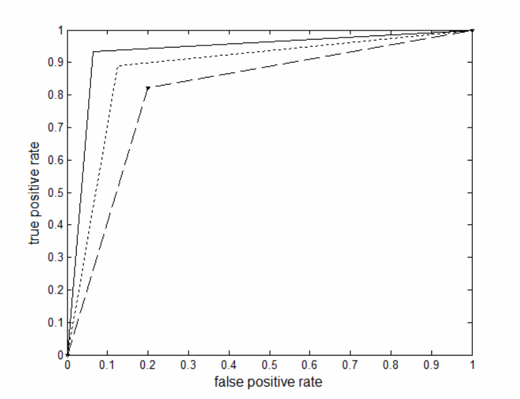


Fig. 4. Averaged ROC curve for testing sets from GLCM (—), GLDM (····), and RLGM (– –) features fed into support vector machine inputs

Although the GLCM method scores the highest AUC value, the GLDM method follows it very closely. The GLRLM method performance is the worst of three but it can still be considered as indicative. These ROC curves confirm the previous results and reinforce the hypothesis of using texture analysis as discriminative information. According to classification results for the three methods we could affirm that texture analysis might offer complementary information to support radiologist on classifying aneurysm evolution after EVAR.

4 Conclusions

The results obtained by each texture analysis method permit to assert that the two main aortic thrombus aneurysm evolutions, namely favourable or unfavourable, correspond to different textures parameters. Consequently, we can conclude that texture analysis could be utilized by physicians as complementary information to classify the post-operative evolution in patients who underwent EVAR treatment. The results can be considered as promising, taking into account the limited number of patients. A bigger patient dataset would be needed in order to generalise the findings to different clinical situations. The study can also be regarded as a first step to more specific studies, particularly for the unfavourable-endotension cases. In these cases a better knowledge of the evolution of aneurysm thrombus by mean of texture analysis could be precious for physicians at the time to decide the treatment to follow.

References

1. Thompson, M.M.: Controlling the expansion of abdominal Aneurysm. Br. J. Surg. 90, 897–898 (2003)
2. VanDamme, Sakalihasan, Limet: Factors promoting rupture of abdominal aortic aneurysms. Acta Chir. Belg. 105(1), 1–11 (2005)
3. William Stavropoulos, S., Charagundla, S.R.: Imaging Techniques for Detection and Management of Endoleaks after Endovascular Aortic Aneurysm Repair. Radiology 243, 641–655 (2007)
4. Bashir, M.R., Ferral, H., Jacobs, C., McCarthy, W., Goldin, M.: Endoleaks After Endovascular Abdominal Aortic Aneurysm Repair: Management Strategies According to CT Findings Am. J. Roentgenol. 192, W178–W186 (2009)
5. Morton, M.J., Whaley, D.H., Brandt, K.R., Amrami, K.: Screening mammograms: interpretation with computer-aided detection-prospective evaluation. Radiology 239, 375–383 (2006)
6. Boniha, L., Kobayashi, E., Castellano, G., Coelho, G., Tinois, E., Cendes, F., et al.: Texture analysis of hippocampal sclerosis. Epilepsia 44, 1546–1550 (2003)
7. Arimura, H., Li, Q., Korogi, Y., Hirai, T., Abe, H., Yamashita, Y., Katsuragawa, S., Ikeda, R., Doi, K.: Automated computerized scheme for detection of unruptured intracranial aneurysms in threedimensional MRA. Acad. Radiol. 11, 1093–1104 (2004)
8. Zhang, J., Tan, T.: Brief review of invariant texture analysis methods. Pattern Recognition 35(3), 735–747 (2002) ISSN 0031-3203
9. Haralick, R., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics SMC 3, 610–621 (1973)
10. Zhang, J., Tan, T.: Brief review of invariant texture analysis methods. Pattern Recognit. 35(3), 735–747 (2002)
11. Weszka, J.S., Dyer, C.R., Rosenfeld, A.: A comparative study of texture measures for terrain classification. IEEE Trans. Syst., Man, Cybern. SMC-6, 269–285 (1976)
12. Mir, A.H., Hanmandlu, M., Tandon, S.N.: Texture analysis of CT images. IEEE Engineering in Medicine and Biology Magazine 14(6), 781–786 (1995)
13. Gibbs, P., Turnbull, L.W.: Textural analysis of contrast-enhanced MR images of the breast. Magn. Reson. Med. 50, 92–98 (2003)
14. Nikita, A., Nikita, K.S., Mougiakakou, S.G., Valavanis, I.K.: Evaluation of texture features in hepatic tissue characterization from non-enhanced CT images. In: 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2007, pp. 3741–3744 (2007)
15. García, G., Maiora, J., Tapia, A., De Blas, M.: Evaluation of texture for classification of abdominal aortic aneurysm after endovascular repair. Accepted for publication in Journal of Digital Imaging
16. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(2), 273–297 (1995)
17. Zhou, X., Wu, X.Y., Mao, K.Z., Tuck, D.P.: Fast Gene Selection for Microarray Data Using SVM-Based Evaluation Criterion. In: IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2008, November 3-5, pp. 386–389 (2008)
18. Ali, A., Khan, U., Tufail, A., Kim, M.: Analyzing Potential of SVM Based Classifiers for Intelligent and Less Invasive Breast Cancer Prognosis. In: 2010 Second International Conference on Computer Engineering and Applications (ICCEA), March 19-21, pp. 313–319 (2010)

Deformable Registration for Geometric Distortion Correction of Diffusion Tensor Imaging

Xu-Feng Yao[1,2] and Zhi-Jian Song[1]

[1] Digital Medical Research Center, Fudan University / The Key Laboratory of MICCAI of Shanghai, Shanghai, 200032, China
zjsong@fudan.edu.cn

[2] Shanghai Medical Instrument College, University of Shanghai for Science and Technology, Shanghai, 200091, China
yao6636329@hotmail.com

Abstract. Geometric distortion of diffusion tensor imaging (DTI) always results in inner brain tissues shift and brain contour deformation and it will certainly lead to the uncertainty of DTI and DTI fiber tracking in the planning of neurosurgeries. In this study, we investigated the accuracy of two deformable registration algorithms for distortion correction of DTI in the application of computer assisted neurosurgery system. The first algorithm utilized cubic B-spline modeled constrained deformation field (BSP) registration of the 3D distorted DTI image to 3D anatomical image, while the second algorithm used multi-resolution B-spline deformable registration. Based on the results, we found that multi-resolution B-spline registration is more reliable than BSP registration for distortion correction of multi-sequence DTI images, the contour deformation and inner brain tissue displacement could be well calibrated in 2D and 3D visualizations. The mesh resolution of B-spline transform plays a great role in distortion correction. This multi-resolution B-spline deformable registration can help to improve the geometric fidelity of DTI and allows correcting fiber tract distortions which is critical for the application of DTI in computer assisted neurosurgery system.

Keywords: Diffusion tensor imaging, Geometric distortion, Deformable registration, B-spline transform.

1 Introduction

Currently, MR DTI is widely utilized in many clinical practices, especially in the field of neurological evaluation and neurosurgical planning [1]. The common scanning pulse sequence of DTI is multi-directional diffusion weighted imaging (DWI) with a fast, single-shot, echo planar imaging (EPI) readout [2]. This fast imaging protocol leads to geometric distortion for the effect of eddy current produced during DTI data acquisition. In the field of image guided neurosurgery, connection of cerebral lesions and adjacent fiber bundles is usually provided by

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 545–553, 2011.

fusing anatomical images and fiber bundles reconstructed by DTI fiber tracking, but the distortion makes the fusion could not provide reliable anatomical relationship. The maximum distortion along the phase-encoding direction can reach 6.5mm at the brain frontal lobe and represents nonlinear characteristics, while distortion along the frequency-coding direction is inconspicuous [3]. Fig. 1 shows the fusion image of an anatomical image and a fiber tract of corpus callosum (CC) reconstructed from raw distorted DTI images. It is easily seen that the CC extends out the brain boundary at the region of frontal lobe (labeled by green arrow).

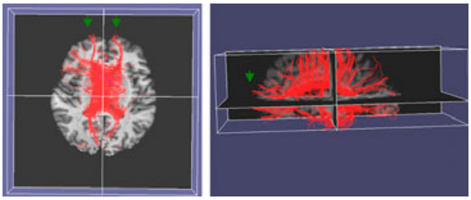


Fig. 1. The fusion of anatomical image and CC tract from raw DTI image

Image registration schemes can be generally classified into two kinds of models for correcting DTI distortion. One kind of these is the linear model. In the work by Mistry et al. [4], the retrospective registration via mutual information (MI) and Fourier transform(FT)-based affine deformations was applied to correct distortion in a 3D high resolution DTI dataset, but the distortions could be not completely corrected. Another kind of these is the nonlinear model. A nonlinear registration using Bezier functions was presented for the correction of susceptibility artifacts in DTI [5]. The comparison of two EPI distortion correction methods in DTI was addressed by Wu et al. [6], BSP showed an overall better performance than B0 field mapping. In summary, the B-spline transform utilized in mentioned nonlinear registrations is effective in distortion correction. Nevertheless, its accuracy still needs further evaluation for severe clinical requirements.

In this study, two deformable registration approaches based on B-spline transform were presented to solve the intractable distortion of DTI. The accuracy and performance of the two proposed method was also compared and addressed in detail in latter sections. This study aims to determine the reliability of using B-spline deformable registration for the system of computer assisted neurosurgery.

2 Materials and Methods

2.1 Image Data

A total of 10 cases of clinical DTI images were provided by the Huashan hospital where the authors are affiliated with. Five cases were acquired from a 1.5T MRI scanner (Signa Twin speed, GE Medical Systems, USA). The other five cases were acquired from a 3T MRI scanner (Signa Excite, GE Medical Systems, USA). DTI was acquired with a multi-slice single shot spin echo (SE) diffusion weighted EPI sequence (TE = 70-90 ms, TR = 7000-8000 ms, slice thickness = 5-7 mm, thick gap = 1 mm, matrix = 128 ×128, FOV = 24 cm, b = 1000 s/mm^2, 6-20 non-collinear gradient directions). The corresponding T1 or T2 weighted image (WI) was acquired with an SE sequence (FOV = 24 cm, TR = 2000-2280 ms, TE = 18-20 ms, slice thickness = 7-10 mm, thick gap = 1 mm). Axial view orientation was used for all images.

2.2 Deformable Registration

The registration has two basic input images. A 3D reference image is defined as the fixed image and a 3D DTI image as the floating image. This process involves an optimization problem to find the spatial map that will align the floating image with the fixed image by iteratively searching the space defined by the transform parameters.

A similarity metric of mutual information (MI) is applied because it is a robust measure criterion for the registration between two images of different modalities [7]. The MI can quantitatively measure how well the transformed floating image fits the fixed image by comparing the gray-scale intensity of the two images. The optimization of Limited memory Broyden Fletcher Goldfarb Shanno with bound constrained (LBFGSB) is used to find the global extreme of the MI criterion due to its efficiency in optimizing high numbers of transform parameters [8]. The Bi-linear interpolator is assisted to estimate the image intensity at the non-grid positions. Here, the 3D deformable transform implemented in deformable registration is either BSP transform or two levels of B-spline transform with different mesh resolutions.

The deformable BSP transform is designed to deal with different scale distortions by setting appropriate 3D mesh resolutions. The transformation model is a 3D free-form deformation (FFD) [9] that can be described by a cubic B-spline as:

$$T_B(x,y,z) = \sum_{l=0}^{3}\sum_{m=0}^{3}\sum_{n=0}^{3} B_l(u)B_m(v)B_n(w)\Phi_{i+l,j+m,k+n} \tag{1}$$

For any point x, y and z of the floating image, the B-spline transform is computed from the positions of the surrounding 4×4×4 control points. The parameter $\Phi_{i,j,k}$ is the set of the deformation coefficients which is defined on a regular lattice of control points placed over the floating image. The spacing between the control points is denoted by δ_x, δ_y and δ_z which represents the lattice mesh resolution. The mesh resolution is inverse proportional to the capability of correcting distortion. The i, j and k are the indices of the control points $i = x/\delta_x$-1, $j = y/\delta_y$-1 and $k = z/\delta_z$-1; u, v and w are the relative positions of (x, y, z) inside that cell in the 3D space:

$$u = \frac{x}{\delta_x} - \left\lfloor \frac{x}{\delta_x} \right\rfloor, v = \frac{y}{\delta_y} - \left\lfloor \frac{y}{\delta_y} \right\rfloor, w = \frac{z}{\delta_w} - \left\lfloor \frac{z}{\delta_w} \right\rfloor \tag{2}$$

The functions B_0 through B_3 are the third-order spline polynomials:

$$\begin{aligned} B_0(t) &= (-t^3 + 3t^2 - 3t + 1)/6 \\ B_1(t) &= (3t^3 - 6t^2 + 4)/6 \\ B_2(t) &= (-3t^3 + 3t^2 + 3t + 1)/6 \\ B_3(t) &= t^3/6 \end{aligned} \tag{3}$$

The two levels of B-spline transform consists of two layers of B-spline transform with appropriate mesh resolutions. It first performed at a coarse level where the two images have fewer pixels and the transform parameters are then used to initialize registration at the next finer level where the two images have more pixels. The whole space transform $T(u, v, w)$ is then defined as:

$$T(u, v, w) = T_{global}(u, v, w) + T_{local}(u, v, w) \tag{4}$$

Where the B-spline transform $T_{global}(u, v, w)$ with low mesh resolution at the first level solves large-scale global distortion, While the B-spline transform $T_{local}(u, v, w)$ with high mesh resolution at the second level is used to solve residual small-scale local distortion after the transform of $T_{global}(u, v, w)$. In this way, the transform parameters of the first level are transferred to the next level for deducing the whole space transform. For every level, the mentioned optimization is used to find the global extreme of transform parameters.

3 Experimental Setup

3.1 Setup of Distortion Correction Method

The distortion correction method used in this study involves three steps. The first step is the 3D brain segmentation of an undistorted reference and a distorted DTI image sequences using brain extraction tool (BET) first proposed by Smith [10]. The aim of 3D brain segmentation is to avoid the disturbance from non-brain tissues to ensure the reliability of the 3D registration. T1WI or T2WI images of SE sequence are taken as the reference images because they explicitly represent anatomical structures and show negligible distortion. The second step is the 3D deformable registration methods mentioned above. The third step is the correction of multiple DTI sequences by the optimal space transform deduced from step two. DTI is a multi-directional imaging modality that includes at least seven sequences for tensor computation, and many other routinely required scalar metric sequences in radiological diagnosis. Through the transform $T\ (u, v, w)$, every 3D floating image can be matched with 3D reference image and its distortion can be rectified.

3.2 Setup of Method Validation

The two registration algorithms with different mesh resolutions are compared for identifying theirs performance. The whole process is accomplished by a panel of three experienced radiologists for the estimation of correction accuracy.

In the 2D visualization, the same 2D physical slices are drawn from 3D images. The maximum brain contour deformation along phase coding direction and the inner brain displacement of lateral ventricle are analyzed by overlying reference image with distorted or corrected DTI images, respectively. Those maximum differences in brain boundary and inner brain displacement are visually inspected and measured with the software of Adobe Photoshop by the radiologists. The measured results are performed with independent samples t test.

In the 3D visualization, the brain contour deformation and inner brain displacement are also compared by visual inspection. During the comparison of inner brain tissue displacement, the CC tracked by fiber assignment by continuous tracking algorithm [11] is fused with corrected or uncorrected fractional anisotropy (FA) images. The match of CC tract and WM in FA images reflect the ability of the two proposed methods in correcting fiber tract distortion.

4 Results

Figs. 2-5 show some examples of the method outcomes. Fig. 2 shows the measurements of maximum brain distortion and inner brain displacement; Fig. 3 shows the comparison of brain contour correction results in 2D; Fig. 4 shows the comparison of inner brain displacements of lateral ventricle in 2D; Fig. 5 shows the comparison of 3D results.

4.1 Visualization of Results in 2D

Figs. 2a and 2b show the measurement of maximum brain contour distortion. The brain boundary of reference image is enclosed by the red rectangle, whereas that of DTI is drawn by the green rectangle. The two overlying images with rectangles is presented in Fig. 2b; The distortion along the phase-coding direction is obvious and it is inconspicuous along the frequency-coding direction. Table 1 records the maximum brain contour distortions along the phase-coding direction without correction. Figs. 2c and 2d display the measurement of inner brain displacement. The lateral ventricle from the reference image is overlaid with the DTI image is presented in Fig. 2d. The anterior horn of the lateral ventricle appears distinct in the non-overlapped regions, whereas no obvious displacement can be found at the posterior horn of lateral ventricle. Table 2 lists the displacements of lateral ventricle without correction.

Fig. 3a shows the overlaid image of reference image and uncorrected DTI image. Fig. 3b shows the overlaid image of reference image and corrected DTI image by the BSP at the mesh resolution of 6×6×5; Fig. 3c shows the overlaid image of reference image and corrected DTI image by the multi-resolution B-spline deformable registration method at the first level with mesh resolution of 6×6×5 and at the second level with mesh resolution of 12×12×5. The brain boundary in the reference image is delineated in green, whereas the brain boundary in the corrected DTI image is drawn in red only at the region of frontal lobe for its obvious distortion. For the two registrations, there still exists some relic distortion

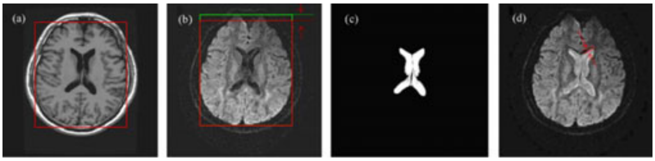


Fig. 2. Measurements of maximum brain contour distortion and inner brain displacement: (a) reference image; (b) DTI image; (c) lateral ventricle image from (a); (d) overlying image of (b) and (c)

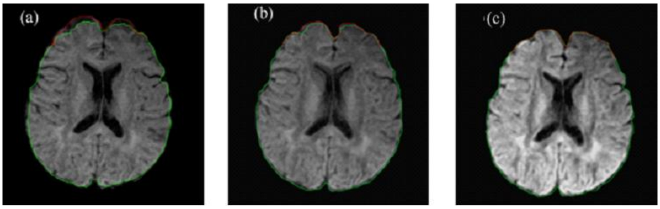


Fig. 3. Comparison of brain contour correction results in 2D: (a) without correction; (b) BSP method; (c) multi-resolution B-spline registration method

after the correction by the BSP and the difference in brain contour is almost negligible by the multi-resolution B-spline method. Table 1 gives the results of the maximum brain distortion along the phase coding direction corrected by the two proposed methods.

Fig. 4a shows the overlaid image of lateral ventricle from reference image and segmented DTI raw image and Fig. 4b shows the corresponding overlaid image of lateral ventricle segmented from the reference image and corrected DTI image. After distortion correction, the displacement of lateral ventricle could be fully rectified without displacement. Table 2 records the displacements of lateral ventricle corrected by the two proposed methods.

Table 1 shows that the maximum distortions in DTI can be corrected by the method of BSP (10.5±1.70 vs. 2.7±0.59, P<0.001, paired t test) and the proposed multi-resolution B-spline registration (10.5±1.70 vs.1.5±0.53 P<0.001, paired t test). The proposed multi-resolution B-spline registration shows better

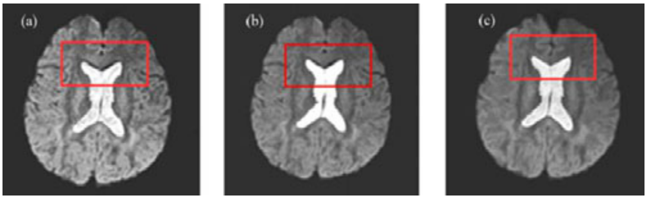


Fig. 4. Comparison of inner brain displacements of lateral ventricle in 2D: (a) without correction; (b) BSP method; (c) Multi-resolution B-spline registration method

Table 1. Comparison of maximum distortion (mm)

Case number	1	2	3	4	5	6	7	8	9	10
Distortions without correction	13.0	10.5	8.5	9.0	11.5	10.0	10.5	13.5	9.5	9.0
BSP correction	2.5	2.5	2.0	2.5	3.0	2.5	3.0	4.0	2.0	3.0
Multi-resolution correction	0.5	1.5	1.5	1.5	2.0	1.5	1.5	2.5	1.0	1.5

performance than the BSP method (2.7±0.59 vs. 1.5±0.53, P<0.001, paired *t* test).

Table 2 shows that the inner displacements in DTI can be corrected by the BSP method (3.6±0.84 vs. 2.35±0.53 P<0.005, paired *t* test) and the proposed multi-resolution B-spline registration method (3.6±0.84 vs. 1.15±0.47 P<0.001, paired *t* test). The proposed multi-resolution B-spline registration method shows better performance than the BSP method (2.35±0.53 vs. 1.15±0.47 P<0.001, paired *t* test).

Table 2. Comparison of inner brain displacements(mm)

Case number	1	2	3	4	5	6	7	8	9	10
Distortions without correction	5.0	3.0	3.0	4.0	3.0	4.0	3.0	5.0	3.0	3.0
BSP correction	3.0	2.0	2.0	3.0	2.5	2.0	2.0	3.0	1.5	2.5
Multi-resolution correction	0.0	1.5	1.0	1.5	1.5	1.0	1.5	1.5	1.0	1.0

4.2 Visualization of Results in 3D

Figs. 5a-5d show the comparison of brain contour in 3D. Fig. 5a shows the 3D reference brain image and Fig. 5b shows the corresponding 3D distorted DTI image. Compared with reference image, contour distortion in the DTI image (indicated by red arrows) is visible at the frontal lobe in Fig. 5b. Fig. 5c shows the corrected DTI image of By BSP and Fig. 5d shows the corrected DTI image by multi-resolution B-spline method. The two proposed methods can minish contour distortion and the multi-resolution B-spline registration is superior to the BSP in brain contour correction; Figs. 5e-5h show the comparison of inner brain displacement of CC in 3D. Fig. 5e shows the overlaid image of the distorted FA and distorted CC and Fig. 5f illustrates the overlaid image of the corrected FA and distorted CC. In Fig. 5f, the reconstructed CC exceeds the region of the FA image and it is inconsistent with the WM in FA image(indicated by green arrows). Figs. 5g and 5h show the overlaid images of corrected FA image and corrected CC. The two proposed method can decrease the inner brain displacement and the multi-resolution B-spline registration is more robust than the BSP in correction of inner brain displacement. It is obvious that both the WM in the FA image and CC are fully matched in Fig. 5h.

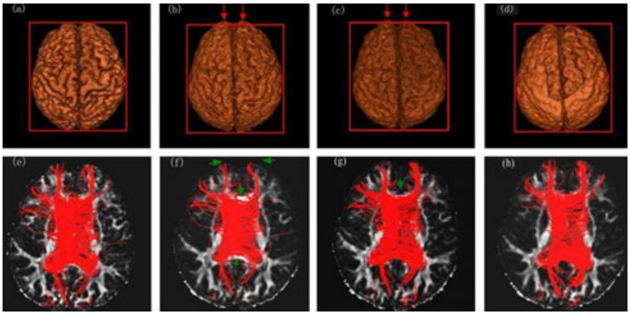


Fig. 5. Comparison of correction results in 3D: (a) 3D referenced image; (b) 3D distorted DTI image; (c) and (d) 3D corrected DTI images; (e) Overlaid image of distorted FA and distorted CC; (f) Overlaid image of corrected FA and distorted CC; (g) and (h) Overlaid images of corrected FA and corrected CC

5 Discussions and Conclusions

The main limitation of DTI in image guided neurosurgery is untruthful anatomical presentations because of the geometric distortion. Besides image processing methods, the distortion correction of DTI can resort to the improvement of hardware and pulse sequence compensation [12]. However, these methods cannot fully calibrate the geometric distortions due to the eddy current, and they are difficult to implement in clinical applications.

In this study, two kinds of B-spline registrations were used to rectify geometric distortions in multiple DTI sequences. multi-resolution deformable registration fully considered the global and local distortions. The mesh resolution of B-spline transformation is a critical for the reason that mesh resolution reflects the space between the control points on the 3D images, whereas an appropriate space can deals with different scale distortion. With the only one B-spline transform, BSP experiences difficulty in solving both global and local distortions only by setting one mesh resolution at one registration.

The results of 2D and 3D displays proved that the multi-resolution B-spline registration not only can correct the outer brain contour but also rectify the inner brain displacement. It can be used as a practical way for distortion correction in the system of computer assisted neurosurgery.

Acknowledgements. The work was supported by the 863Program of China (No. 009AA02Z415), the NNS Fund of China (Nos. 30570506 and 60972102), the Science and Technology Commission of the Shanghai Municipality of China (Nos. 08dz1900705, 09410702800 and 10dz2211800), the SLAD Project (No. B112) and the IP of SMEC (No. 11yz292).

References

1. Chanraud, S., Zahr, N., Sullivan, E.V., Pfefferbaum, A.: MR Diffusion Tensor Imaging: A Window into White Matter Integrity of the Working Brain. Neuropsychol. Rev. 20, 209–225 (2010)
2. Le Bihan, D., Poupon, C., Amadon, A., Lethimonnier, F.: Artifacts and pitfalls in diffusion MRI. J. Magn. Reson. Imaging 24, 478–488 (2006)
3. Mattila, S., Renvall, V., Hiltunen, J., Kirven, D., Sepponen, R., Hari, R., Tarkiainen, A.: Phantom-based evaluation of geometric distortions in functional magnetic resonance and diffusion tensor imaging. Magn. Reson. Med. 57, 754–763 (2007)
4. Mistry, N.N., Hsu, E.W.: Retrospective distortion correction for 3D MR diffusion tensor microscopy using mutual information and Fourier deformations. Magn. Reson. Med. 56, 310–316 (2006)
5. Merhof, D., Soza, G., Stadbauer, A., Greiner, G., Nimsky, C.: Correction of susceptibility artifacts in diffusion tensor data using non-linear registration. Med. Image Anal. 11, 588–603 (2007)
6. Wu, M., Chang, L.C., Walker, L., Lemaitre, H., Barnett, A.S., Marenco, S., Pierpaoli, C.: Comparison of EPI distortion correction methods in diffusion tensor MRI using a novel framework. Med. Image Comput. Comput. Assist. Interv. 11, 321–329 (2008)
7. Khader, M., Ben Hamza, A., Bhattacharya, P.: Multimodality Image Alignment Using Information-Theoretic Approach. In: Campilho, A., Kamel, M. (eds.) ICIAR 2010. LNCS, vol. 6112, pp. 30–39. Springer, Heidelberg (2010)
8. Zhu, C.Y., Byrd, R.H., Lu, P.H., Nocedal, J.: Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. Acm T. Math. Software 23, 550–560 (1997)
9. Mattes, D., Haynor, D.R., Vesselle, H., Lewellen, T.K., Eubank, W.: PET-CT image registration in the chest using free-form deformations. Ieee T. Med. Imaging 22, 120–128 (2003)
10. Smith, S.M.: Fast robust automated brain extraction. Hum. Brain Mapp. 17, 143–155 (2002)
11. Stadlbauer, A., Nimsky, C., Buslei, R., Salomonowitz, E., Hammen, T., Buchfelder, M., Moser, E., Ernst-Stecken, A., Ganslandt, O.: Diffusion tensor imaging and optimized fiber tracking in glioma patients: Histopathologic evaluation of tumor-invaded white matter structures. Neuroimage 34, 949–956 (2007)
12. Nana, R., Zhao, T.J., Hu, X.P.: Single-Shot Multiecho Parallel Echo-Planar Imaging (EPI) for Diffusion Tensor Imaging (DTI) With Improved Signal-to-Noise Ratio (SNR) and Reduced Distortion. Magn. Reson. Med. 60, 1512–1517 (2008)

Automatic Localization and Quantification of Intracranial Aneurysms

Sahar Hassan[1,2], Franck Hétroy[1,2], François Faure[1,2], and Olivier Palombi[1,2,3]

[1] Université de Grenoble & CNRS, Laboratoire Jean Kuntzmann, Grenoble, France
[2] INRIA Grenoble - Rhône-Alpes, Grenoble, France
[3] Grenoble University Hospital, France

Abstract. We discuss in this paper the problem of localizing and quantifying intracranial aneurysms. Assuming that the segmentation of medical images is done, and that a 3D representation of the vascular tree is available, we present a new automatic algorithm to extract vessels centerlines. Aneurysms are then automatically detected by studying variations of vessels diameters. Once an aneurysm is detected, we give measures that are important to decide its treatment. The name of the aneurysm-carrying vessel is computed using an inexact graph matching technique. The proposed approach is evaluated on segmented real images issued from Magnetic Resonance Angiography (MRA) and CT scan.

1 Introduction

Aneurysms are dilatations in the wall of a blood vessel, leading to little pockets. Aneurysms can be saccular, fusiform or dissecting, see Fig. 1. In this article we are interested in saccular aneurysms which are connected to the vessel by a narrowed zone called the *neck*. If not treated, an aneurysm may burst causing a stroke and in most cases the death of the patient.

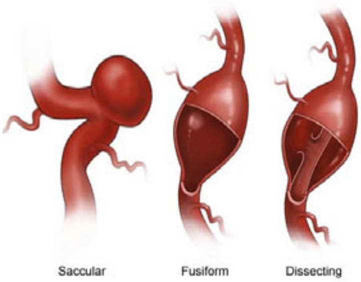


Fig. 1. Aneurysm types[1]

The decision of treating an aneurysm or just observing it is made according to its risk of rupture. When the treatment is needed, two possible ways exist: either embolization using a platinum coil, or clipping. A lot of studies and statistical surveys have been done in order to know what factors affect the rupture of an aneurysm [1,2,3], and thus help in making the best decision about the treatment. According to these studies the most important factors are: size, shape, neck, and location of the aneurysm.

A lot of work has been done in the domain of intracranial aneurysms, most of which is about segmenting the vascular tree and giving the user a 3D view of the aneurysm. This segmentation can be statistical [4], or it can be based on the tubular shape of vessels [5,6,7,8]. In [9], a morphological characterization of the aneurysm is given in order to predict the rupture rate, and thus decide if there should be a treatment.

[1] http://nyp.org/health/neuro-cerbaneu.html

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 554–562, 2011.

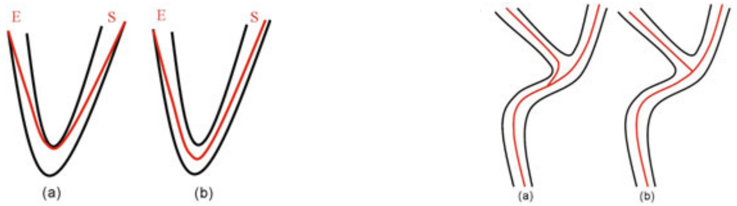
In this paper, we suppose that the segmentation is done, and we go further. The set of voxels representing the cerebrovascular tree goes through several processes including: extraction of vessels' centerlines, detection of aneurysms, quantification and localization of the detected aneurysms. An approach based on Dijkstra's algorithm [10] is proposed to get thin, connected and centered centerlines. These centerlines are then used to study the evolution of the diameters and automatically detect aneurysms. Blood vessels have a cylindrical shape and thus their diameters are almost steady, whereas those of aneurysms change considerably. Relevant measures of found aneurysms and their location are then given using a partial graph matching technique. To our knowledge, this is the first time these steps are performed together to detect, quantify and localize intracranial aneurysms.

2 Methods

2.1 Centerlines Extraction

Extraction of blood vessels centerlines can be done either while segmenting these blood vessels [7,8,11,12], or after segmenting blood vessels from medical images as in our case. Various methods for centerline extraction are proposed for different uses. Some categories of these methods are presented in [13] along with the usually desired properties of centerlines. Since we want to use the centerlines to study the evolution of blood vessels diameters, these centerlines should be: **1. connected**: the centerlines we are looking for should be 26-connected, **2. thin**: a centerline is thin if each voxel of the centerline has only two of its neighbors in the centerline, except for the extremities which have one neighbor in the centerline, **3. centered**: the centerlines should be centered within the vascular tree, and **4. connections between branches**: should be as perpendicular as possible, see Fig. 2. Finally, the algorithm should be efficient since it is a step out of four in the processing chain, besides cerebral vascular trees are complex.

In the following, we call *skeleton* the set of centerlines. The longest centerline is called the *main centerline*, while the others are called *branches*. The main centerline and each branch have a diameter which is the mean diameter of the corresponding blood vessels.

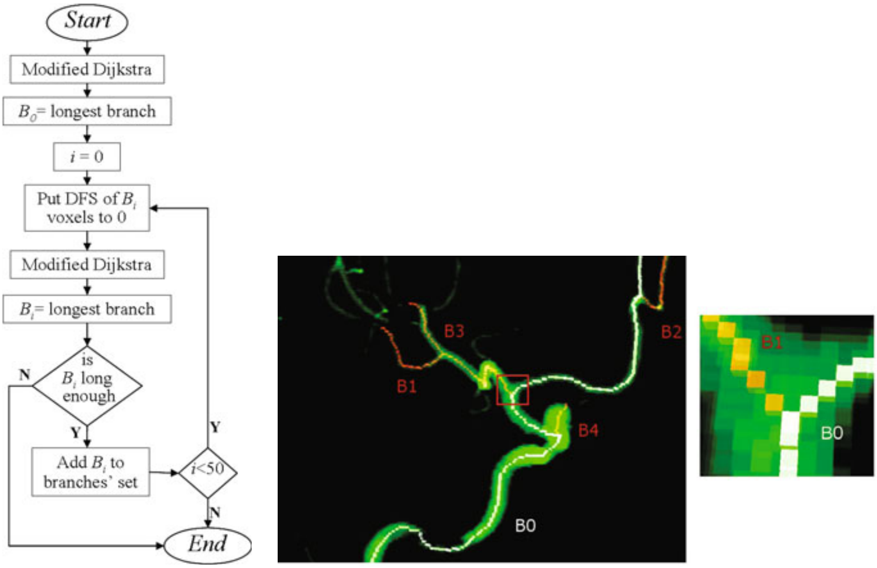


Main centerline using Dijkstra's algorithm with Euclidian distance (a), the wanted centered centerline (b)

Connection between branches: (a) the connection is not perpendicular, (b) the wanted perpendicular connection

Fig. 2. Important features of the desired centerlines

To fulfil our requirements, we propose a centerline extraction method that falls in the **distance-based methods** category. The main idea of these methods is to construct a shortest distance tree (SDT) [10]. After the construction of such a tree, we get a graph. Nodes of the graph are the voxels of the object. The voxels will be connected (a connection between two voxels corresponds to an edge in the graph) in a way to minimize the distance to a source voxel S, hereafter called Distance From Source (DFS). The main centerline is then extracted by tracing from E, the voxel with maximum DFS, back to the source S, and thus is connected and thin by construction. The use of a heap for the priority queue makes the complexity of these methods of $O(NlogN)$ where N is the number of voxels and thus computationally efficient. However, using the Euclidian metric as the distance to minimize leads to a centerline that cuts the corners, see Fig. 2. Several variations of this algorithm were proposed to solve the "cutting corners" problem and get centered centerlines [14,15,16]. The common idea is to use another distance function while constructing the tree to privilege voxels near the center of the object.



(a) (b)

Fig. 3. Our method. (a) Flowchart. 50 is a sufficient number to extract all significant branches in all our experiments. (b) Result of the method on a real dataset with a zoom on the branching.

Our method is illustrated on Fig. 3. **First**, the source voxel is chosen automatically, to be sure that it is an extremity of a vessel. We construct a SDT taking an arbitrary voxel as a source, the end voxel (furthest one of the arbitrary source) is necessarily an extremity and is used as the source voxel for our algorithm.

We use the following distance function instead of using the Euclidian distance:

$$d(v1, v2) = \frac{dist(v1, v2)}{1 + (DFB[v1] + DFB[v2])}$$

with: $dist(v1, v2)$ the Euclidian distance between $v1, v2$, $DFB[v_i]$ v_i's distance from the boundary, i.e. Euclidian distance between v_i and the closest surface voxel (a surface voxel is a voxel with at least one of its 26-neighbors missing in the voxel set).

The division by the distance from boundary (depth) privileges the voxels that are far from the boundary, and thus enforce the centeredness. At the same time, we keep using the Euclidian distance to find the *end voxel* at each iteration, and thus extract branches in a descending length order. Each branch B_i is connected to a *father* branch that is not necessarily B_{i-1}. Another important advantage of our algorithm is junctions between branches. Putting DFS of voxels of extracted branches to zero, makes each branch join its *father* in a perpendicular way (see Fig. 3). We emphasize on this point because variations of branches' diameters play a major role in aneurysm detection and quantification, see Section 2.2.

The complexity of our algorithm is $O(KNlogN)$ where K is the number of extracted branches, and N is the number of voxels. One drawback of this method is that the set of branches is not homotopic to the object. This method gives by construction a tree-like structure with no loops.

2.2 Automatic Detection of the Aneurysm

One key characteristic that differentiates a saccular aneurysm from a normal vessel, is that the normal vessel -which has a cylindrical shape- has an almost steady diameter, whereas the aneurysm -which has an irregular shape- has a diameter that changes considerably.

In order to model the appearance of a vessel, we define a set of points (x, y). Each point corresponds to a voxel v of the branch, where:

• x, represents the distance between the voxel v and the origin of the branch j.

• y, represents the approximate diameter of the branch at v.

To calculate y, we compute the real plane P passing through the center of voxel v and perpendicular to the branch, see Fig. 4. P cuts the vessel or aneurysm surface on voxels $v_i, 1 \leq i \leq k$. Let y_i be the distance between v_i and v, y is defined as the average value of $y_i : y = \frac{\sum_{i=1}^{k} y_i}{k}$. Thanks to the centeredness of centerlines, and perpendicular connections between branches, y represents a reliable measure of the diameter changes.

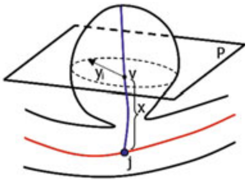
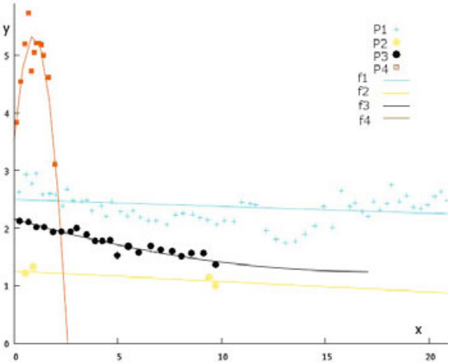


Fig. 4. Calculating y

Then, we use the least-squares method to find the quadratic function ($y = a + bx + cx^2$) that best matches our set of points. A more complex function could be used, but this one is sufficient to discriminate between a diameter variation which is linear and a one that is not. Since normal vessels have a cylindrical shape, their diameter is almost steady and thus the value of c is very small. So, by thresholding on c, we decide if the corresponding branch is in an aneurysm. The threshold we use has been found after a ROC analysis, and is 0.2. The threshold is not null because a branch can traverse several blood vessels (see branches in Fig. 3), which makes the associated diameter change. However, this change remains insignificant in comparison with the one caused by an aneurysm.



(a)

Branch	a	b	c
B_1	2.494	-0.012	0.000
B_2	1.250	-0.018	0.000
B_3	2.156	-0.105	0.003
B_4	3.509	3.831	-2.006

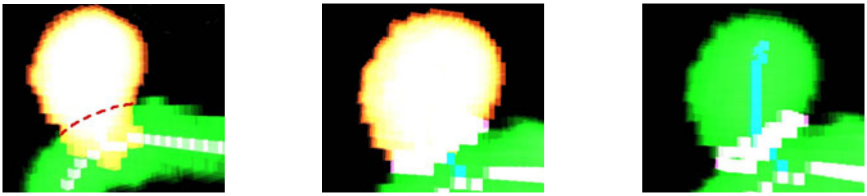
(b)

Fig. 5. Diameters variations for branches of the real dataset shown in Fig. 3: (a) The quadratic functions, note that they closely match straight lines for vessels, which is not the case for the one of the aneurysm (B_4). (b) Table1 shows values of a,b and c for each branch.

During the extraction of branches, the above test is made on each branch B_i to decide if it is an aneurysm or not. Branches that are in aneurysms are saved in a list to be treated later for quantification.

2.3 Aneurysm Quantification

The construction of a shortest distance tree creates an oriented graph. The nodes of the graph are the voxels. The oriented edges link these voxels together to minimize their distance from the source voxel. Voxels of an aneurysm are the voxels that can be reached from voxels of the *aneurysm branch* by descending the graph. Since the *aneurysm branch* is connected to the *father* branch, which is inside the holding vessel, some of its voxels are inside the holding vessel, see Fig. 6-(a). In order to get rid of these voxels, we only add voxels if their distance from the branch of the holding vessel is greater than its radius, see Fig. 6-(b). The aneurysm's neck is the surface voxels of the aneurysm that have at least one neighbor that is not in the aneurysm, see Fig. 6-(c),(d).



(a)In yellow, voxels linked to those of the aneurysmal branch.

(b) The voxels of the aneurysm.

(c) The neck of the aneurysm.

Fig. 6. Compute aneurysm's neck

Following a discussion with a surgeon, we found out that the following measures of the aneurysm are relevant to help the treatment decision:

- Size of the aneurysm: number of aneurysmal voxels.
- Maximum vertical diameter of the aneurysm ($Diam1$): to find this diameter, we look for the surface voxel which is the furthest from the origin j of the aneurysmal branch. $Diam1$ is the distance between this voxel and j.
- Maximum horizontal diameter of the aneurysm ($Diam2$): we look for the voxel m of the aneurysmal branch with maximum DFB, then $Diam2 = 2 \times DFB[m]$.

2.4 Localization of the Aneurysm

Regarding the method we use to extract centerlines, the result is a set of branches where each branch B_i (except B_0) has a father branch. On the same time, the branches do not correspond to blood vessels, a branch can be within several blood vessels. To get a graph that represents the resulting tree, we deal with segments. A segment is made of the voxels of a branch between its extremity and a junction, or between two successive junctions. We choose the widest segment (aneurysms excluded) as root, because it corresponds to the carotid (widest blood vessel), and we construct a graph. In Fig. 8-(a), we see the graph corresponding to the dataset of Fig. 5-(a).

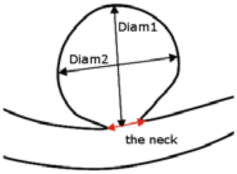
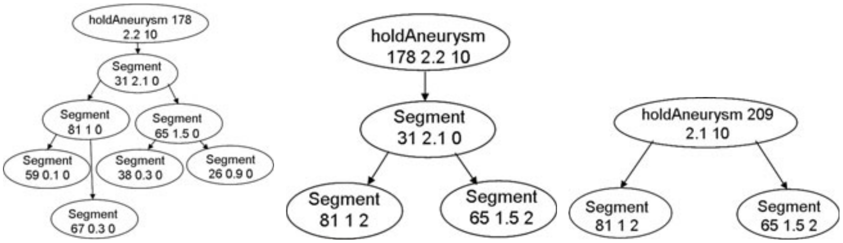


Fig. 7. Measures of an aneurysm



(a)Initial graph (b)Graph without small branch (c)Final graph

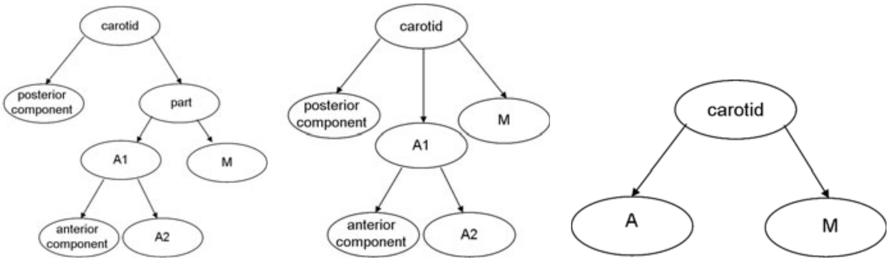
Fig. 8. Graphs for the dataset of Figure5-(a)

Graph matching is a well known problem, and graphs can be with or without attributes for both nodes and edges. If we consider our graph of segments without any attributes, the matching process will be mainly a topological one, meaning that if a node has two child nodes, it may be matched with any node with two children in the reference graph. To get a more accurate matching, we choose to use a graph with attributes.

As can be seen in Fig. 8-(a), we associate to each node of the graph three attributes: length, diameter of the segment, and number of children. The first two attributes are used to give an idea about the importance of the segment. Segments with small diameters or short lengths are considered very patient specific and unimportant. The corresponding nodes are then deleted from the graph (Fig. 8-(b)). We can describe this deletion step as a simplification of the graph. To keep a trace of the deleted nodes, we use the third attribute "number of children". Each time we decide to delete a node, we increase the number of children of its parent by one. Finally, we give the root of our graph a big number of children (10), to be sure that the root will be matched with the carotid.

Only the third attribute (number of children), is then used in the matching step. It helps to differentiate between vessels that are known to have a lot of bifurcations (vessel M) and those who have less bifurcations (vessel A), and both issued from the same parent (carotid), see Fig. 9.

Since the anatomy of the cerebral vascular tree is known, especially regarding the main vessels, we use a reference graph. In practice, not all vessels are segmented from acquired images, so several reference graphs with different resolutions are needed. Fig. 9 shows the reference graphs we use.



(a) (b) (c)

Fig. 9. Reference graphs

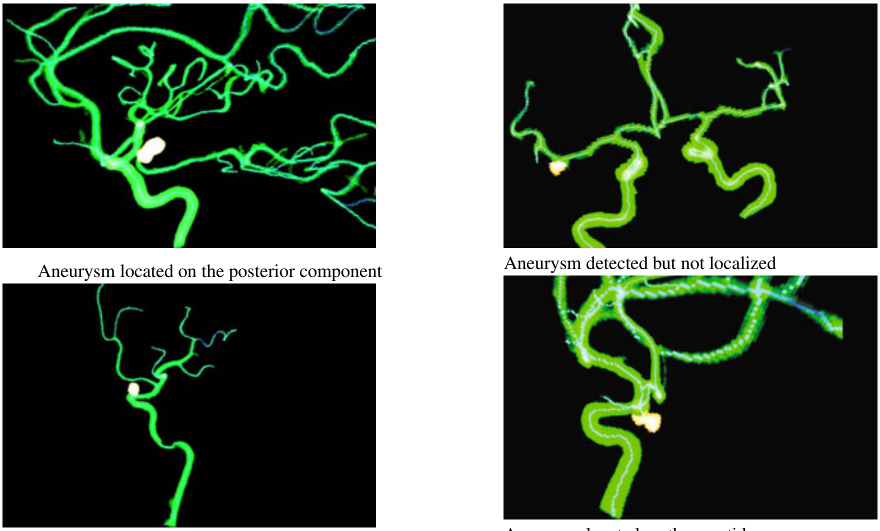
The localization of the aneurysm is then reduced to an inexact graph matching problem. We use the VF algorithm [17] to solve this problem. We try first to match our simplified graph with the most detailed reference graph 9-(a), then with 9-(b), and finally with 9-(c). In practice, more reference graphs can be used if needed.

3 Results

We validated our approach on a set of twenty patients, using both MRA and CT imaging techniques for five and fifteen patients respectively. The set contained five males and fifteen females, the patients' ages varied from 33 to 78 years with an average of 51.68.

After segmentation, our method is applied on one connected component (either chosen by the user, or the largest one if no choice is made). The results reported no error of typeI (false negative) and two errors of typeII (false positive). Results of quantifications were compared to those provided by experts (experts provided quantifications for only 10 cases). We use the following formule to calculate the error of a measurement: $E = 100 \times \frac{\|provided - calculated\|}{provided}$. For $Diam1$, the error varied from 0.8 to 48 with an average of 11.7, for $Diam2$, it varied from 1.7 to 17.1 with an average of 8.25.

Since our technique of localization does not consider cases where the whole cerebrovascular tree is present, the localisation was possible in ten cases and the localizations were distributed as follows: six aneurysms were localized on the carotid, two on



Aneurysm located on the posterior component

Aneurysm detected but not localized

Aneurysm located on the carotid

Aneurysm located on the carotid

Fig. 10. Some examples of aneurysms detected by our method

vessel $A1$ and two on the posterior component. Fig. 10 shows some examples of the detected aneurysms. Calculation time on a Pentium(R) 4 CPU 3.00 GHz varied from 4.5 to 145.23 seconds with an average of 29.97. In practice, this time is almost linearly connected to the number of voxels.

4 Conclusion and Future Work

In this paper, we have presented a complete solution to automatically localize and quantify intracranial saccular aneurysms. First, we use a new distance-based method to find centerlines of the vascular tree. The centerlines are connected, thin (by construction), and centered, due to our modification of Dijkstra's algorithm. Moreover, since the distance map is calculated relative to a source voxel, the presented approach is invariant to rigid transformations. Then, aneurysms are automatically detected and quantified. Finally, the aneurysm is localized by graph-subgraph matching between a graph representing the centerlines and a reference graph.

When applying our method to 3D medical images, it proved to be fast and robust since the quality of the results is independent of small segmentation artifacts.

References

1. Ujiie, H., Tachibana, H., Hiramatsu, O., Hazel, A., Matsumoto, T., Ogasawara, Y., Nakajima, H., Hori, T., Takakura, K., Kajiya, F.: Effectes of size and shape (aspect ratio) on the heomdynamics of saccular aneurysms: a possible index for surgical treatment of intracranial aneurysms. Neurosurgery 45, 119–130 (1999)

2. Weir, B.: Unruptured intracranial aneurysms: a review. J. Neurosurgery 96, 3–42 (2002)
3. Ecker, R., Hopkins, L.: Natural history of unruptured intracranial aneurysms. Neurosurg Focus 17(5) (2004)
4. Wilson, D.L., Noble, J.A.: Segmentation of cerebral vessels and aneurysms from mr angiography data. In: Duncan, J.S., Gindi, G. (eds.) IPMI 1997. LNCS, vol. 1230, pp. 423–428. Springer, Heidelberg (1997)
5. Aylward, S., Pizer, S., Eberly, D., Bullitt, E.: Intensity ridge and widths for tubular object segmentation and description. In: IEEE Workshop on Mathematical Methods in Biomedical Image Analysis, p. 0131 (1996)
6. Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A.: Multiscale vessel enhancement filtering. In: Wells, W.M., Colchester, A.C.F., Delp, S.L. (eds.) MICCAI 1998. LNCS, vol. 1496, pp. 130–137. Springer, Heidelberg (1998)
7. Wink, O., Niessen, W., Viergever, M.: Multiscale vessel tracking. Medical Image Analysis 23(1), 130–133 (2004)
8. Descoteaux, M., Collins, D.L., Siddiqi, K.: A geometric flow for segmenting vasculature in proton-density weighted mri. Medical Image Analysis 12(4), 497–513 (2008)
9. Millán, R.D., Dempere-Marco, L., Pozo, J., Cebral, J., Frangi, A.: Morphological characterization of intracranial aneurysms using 3-d moment invariants. IEEE Transactions on Medical Imaging 26(9), 1270–1282 (2007)
10. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numerische Mathematik 1, 269–271 (1959)
11. Aylward, S.R., Bullitt, E.: Initialization, noise, singularities, and scale in height ridge traversal for tubular object centerline extraction. IEEE Transactions on Medical Imaging 21(2), 61–75 (2002)
12. Deschamps, T., Cohen, L.: Fast extraction of minimal paths in 3D images and applications to virtual endoscopy. Medical Image Analysis 5(4) (2001)
13. Cornea, N., Silver, D., Min, P.: Curve-skeleton properties, applications, and algorithms. IEEE Transactions on Visualization and Computer Graphics 13(3), 530–548 (2007)
14. Bitter, I., Sato, M., Bender, M., McDonnell, K.T., Kaufman, A., Wan, M.: CEASAR: a smooth, accurate and robust centerline extraction algorithm. In: VIS 2000: Proceedings of the Conference on Visualization 2000, pp. 45–52. IEEE Computer Society Press, Los Alamitos (2000)
15. Bitter, I., Kaufman, A.E., Sato, M.: Penalized-distance volumetric skeleton algorithm. IEEE Transactions on Visualization and Computer Graphics 7(3), 195–206 (2001)
16. Wan, M., Liang, Z., Ke, Q., Hong, L., Bitter, I., Kaufman, A.E.: Automatic centerline extraction for virtual colonoscopy. IEEE Trans. Med. Imaging 21, 1450–1460 (2002)
17. Cordella, L.P., Foggia, P., Sansone, C., Vento, M.: An improved algorithm for matching large graphs. In: 3rd IAPRTC15 Workshop on Graph-based representations in Pattern Recognition, pp. 149–159 (2001)

A New Ensemble-Based Cascaded Framework for Multiclass Training with Simple Weak Learners

Teo Susnjak, Andre Barczak, Napoleon Reyes, and Ken Hawick

Massey University Albany, New Zealand
{T.Susnjak,a.l.barczak,n.h.reyes,k.a.hawick}@massey.ac.nz

Abstract. We present a novel approach to multiclass learning using an ensemble-based cascaded learning framework. By implementing a multiclass cascaded classifier with AdaBoost, we show how detection runtimes are accelerated since only a subset of the ensemble is executed, thus making the classifiers suitable for computer vision applications. We also propose a new multiclass weak learner and demonstrate the framework's ability to achieve arbitrarily low training errors in conjunction with it. We tested our algorithm against AdaBoost.OC, ECC and M2 multiclass learning methods, on seven benchmark UCI datasets. In our experiments, we found that our framework achieves higher accuracy on five out of seven datasets and displays faster runtime efficiency in all cases.

1 Introduction

Many real-world classification problems involve predictions that require an assignment to one of multiple classes. Since the probability of making a wrong prediction in multiclass problems is higher than for binary classification, multiclass problems are considered inherently more difficult to solve especially as class numbers increase [1]. Additionally, developing object detection systems that are not only robust, but also real-time capable is an important goal of the computer vision community [2].

A large body of research has shown the effectiveness of boosting and ensemble-based learning to provide robust and efficient solutions to binary class problems; nevertheless, multiclass domains still present a formidable challenge to current approaches. Most methodologies attempt to solve multiclass classification by reformulating the task into a series of binary class problems [3] that results in multiple classifiers being created. The most popular of these are one-against-all (OAA) and one-against-one (OAO) training approaches [4,1], which carry a high runtime cost since all classifiers require execution before a classification is possible.

Effective and theoretically proven extensions of AdaBoost have been proposed for multiclass problems. Of these, AdaBoost.M2 proceeds by implementing a weak learner that selects a set of *plausible* classes for a given sample at each iteration and evaluates each hypothesis based on the related *pseudoloss* measure. Other approaches take advantage of principles behind error-correcting output codes (ECOC) [5]. The best known ECOC-based methods AdaBoost.ECC (error-correcting code) [6] and AdaBoost.OC (output code) [7], iteratively construct coding matrices with uncorrelated errors at training, that become vital for accurate classification resolution.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 563–570, 2011.

Although the outputs of multiclass AdaBoost algorithms are single classifiers, the entire ensemble must still be executed first, before a prediction can be made. This may not satisfy real-time demands of high-speed data streams associated with vision detection. In addition, multiclass AdaBoost often cannot rapidly converge to adequately low training errors on difficult datasets when using weak learners, which results in larger and computationally costlier ensembles. In such instances, more sophisticated learners like C4.5 and CART are employed [8]; however, this trade-off leads to protracted training runtimes.

In this paper we propose a novel multiclass learning method that decomposes the training and detection task into cascades. We present also a new weak learner and demonstrate how it can be combined with the cascaded architecture to attain arbitrarily low training errors and accurate classifiers compared to current multiclass AdaBoost approaches. We address the problem of detection runtimes for real-time critical domains and show how our cascaded classifier need only execute a subset of its collective ensemble in order to formulate a prediction, making it suitable for computer vision applications. [2] have already aptly demonstrated how a multiclass problem can be decomposed into cascades for multi-view face detection within the context of rare-event detection. We put forward a method that is applicable to other, more general problems also.

The suceeding section of the paper describes the weak learner and the details of our cascaded multiclass architecture. We tested our classifiers on seven benchmark multiclass University of California at Irvine (UCI)[1] datasets, whose results we discuss in the remainder of the paper.

2 Multiclass Cascade Learning

We begin first by describing the details of our weak learning algorithm and boosting before outlining the structure of the cascaded multiclass framework.

Inspired by the speed and efficiency of calculating optimal thresholds for simple decision stump learners on binary class problems as demonstrated by [9], we extended this underlying principle to calculating multiple optimal thresholds for *k*-class problems. In effect, we end up with a domain partitioning weak learner that is an extension of [10], in that that partitions are not necessarily disjointed. Given a sorted vector of *k*-class feature values, we first calculate the optimal threshold value and direction for each class label. Our learner achieves this by manipulating the weight distribution of a current example at each step of the traversal *k*-times with respect to all class labels. The manipulation of the weight distribution of a current example is a form of normalization. The effect is that for each class label the weight of the current example is dynamically altered to reflect a binary class distribution in which the allocation of weights is 50-50 between the target class label and the rest.

The first step of calculating optimal *binary*-class distribution thresholds for a vector of *k*-class feature values is not sufficient on its own since it does not generate enough discriminatory information that is necessary to resolve conflicting predictions of multiple class labels for a given sample. We counter this problem by calculating a *secondary*

[1] URL ”http://archive.ics.uci.edu/ml/”

Algorithm 1. Cascaded multiclass learning

```
**INPUT**: training examples $(v_1, y_1), (v_2, y_2), ...(v_n, y_n)$ where $v_i$ is a feature vector $\in V$
and $y$ is a class label
**OUTPUT**: multiclass weak classifier ($t_k^f$,$v^f$, $me_k$) $k$ = class, $K$ = total number of
classes, $w_i$ = weight for an element $i$, $v_i^f$ = value of feature $f$ for element $i$, $t_k^f$ =
threshold/direction on feature $f$, $S_k^+$ = total sum of weights for class $k$, $S_k^-$ = total sum of
weights for non-class $k$, $W_k^+$, $W_k^-$ = current sum of weights, $nc_k$ = normalization
coefficient for class $k$, $error_k^{right/left}$= classification errors on each direction, $me_k$ =
minimum classification error for a certain class $k$,
  **for** each hypothesis **do**
    compute $nc_k$
    initialise $S_k^+$, $S_k^-$
    **for** each feature $f$ **do**
      sort $v^f$
      **for** each $i$ in $v^f$ and each class $k...K$ **do**
        $W_k^+ = \sum(nc_k.w_i)$, **where** $y_i = k$
        $W_k^- = \sum(nc_k.w_i)$, **where** $y_i \neq k$
        $error_k^{right} = W_k^+ + S_k^- - W_k^-$
        $error_k^{left} = W_k^- + S_k^+ - W_k^+$
        $me_k = MIN(error_k^{right}, error_k^{left}, me_k)$
        **if** (new minimum $me_k$) **then**
          store ($t_k^f$,$v^f$, $me_k$)
  repeat algorithm for secondary thresholds in respect to the corresponding primary $t_k^f$
```

threshold on the same feature vector which complements the original *primary* threshold. The role of the secondary threshold is to define an optimal value at which the primary threshold is bound with respect to its direction. The result of this is that the feature vector is partitioned into *k* bins or intervals with every interval representing a class label (Figure 1). Each partition is assigned a confidence value based on its accuracy and the average error rate of all partitions defines the overall error rate for the given feature.

We use binary AdaBoost in order to re-weight all samples after generating each hypothesis. The weight of incorrectly classified samples is increased in proportion to the competence of the last hypothesis and in the case of a sample value falling into a region of overlapping class partitions, a prediction is awarded to a class label associated with the partition carrying the highest confidence.

The multiclass cascade we propose consists of *k* number of layers with each layer trained to predict a given class label. The cascade is also two dimensional whereby each layer contains within it a further *nested* cascade to facilitate the training process and ensure low training error (Figure 1). For clarity we will refer to each layer of a *nested* cascade denoted as M_n in this figure, as a *node*.

The training proceeds as follows: the initial node of the first layer is trained on all samples until a predefined number of boosting iterations are completed. Once this criterion is met, the node is assessed for accuracy and the *correctly* predicted samples belonging to the best performing class label are removed from further training of the current layer. The training for the subsequent node proceeds with samples belonging to

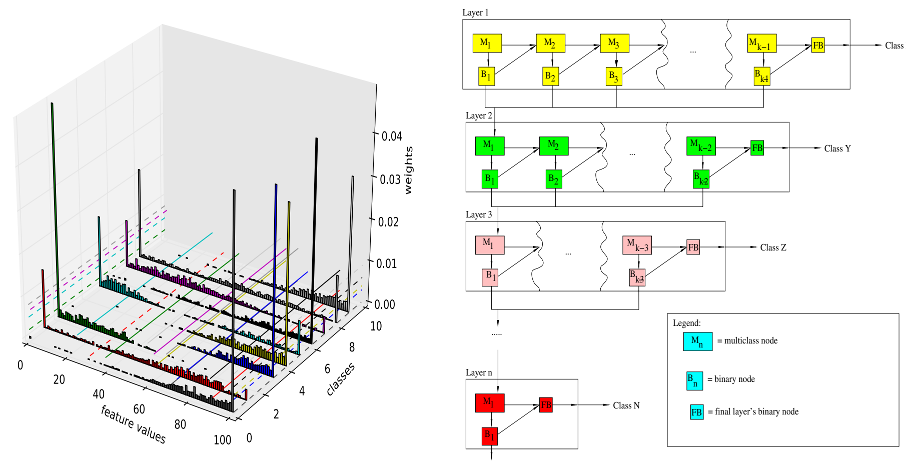


Fig. 1. Example of our weak learning algorithm partitioning the feature vector space on the Pendigits UCI dataset (left). Diagram of the architecture of the proposed cascaded multiclass framework (right).

k-1 class labels until the training essentially becomes a binary problem, after which the class label with the best accuracy becomes the designated label for the layer. Consequently, as each new node is constructed, training takes place on samples of class labels that are most difficult to discriminate from one another while the easier samples are removed in order to facilitate the process.

After a layer has been completed, cascade training restarts in the same fashion, only this time without samples belonging to the class label that was designated to the previous layer. As a result, the training problem is continuously decomposed into simpler learning tasks containing a total number of *k*-1 class labels until a layer for each class label has been trained. The final cascade structure with all its constituent nodes can be visualized as an inverted pyramid.

The selection of the most *appropriate* class label with its corresponding samples for removal after each node is trained, is a critical component of this learning framework. The hit rate of the selected class determines the proportion of its samples that will be removed from further training of the layer. If this is low then the ongoing training of the given layer will not benefit by becoming appreciatively easier for the task of discriminating between classes. Also the training will not become less computationally expensive and most importantly the misclassified samples of a given node will no longer have the possibility of being correctly learned, thus ensuring that the training error will increase.

The false positive rate of the selected class is crucial since these samples may comprise instances of the class label that is eventually to be designated to the layer which we do not know *a priori*. If that is the case, then an arbitrarily low training error will not be attained and the given layer will not generalize well. Therefore the ideal scenario is that a selected class label achieves near 100% hit rates while attaining the vital false positive rates at 0%. Since this is not likely to occur consistently on majority of difficult datasets using the simple weak learner described earlier, a solution is required to handle the false positive predictions after each multiclass node is trained.

We formulate the solution to the problem of false positive detections by training an additional *auxiliary node* attached to the original with the major difference that the supporting node is trained as a binary class problem using simple binary thresholds. The training set for the auxiliary nodes is comprised of the correctly predicted samples belonging to the selected class label of the multiclass node which form the *positives* while the *negatives* consist of all samples that are the false positives. Boosting iterations are executed in the auxiliary node until all the negatives have been correctly classified. The negatives are returned to the layer where the multiclass training continues while the correctly classified positives are removed and forwarded to the subsequent layer for training.

The problem of how to separate remaining samples from the final class after the last multiclass node has been trained is addressed using the same strategy in order to ensure zero training error for each layer. The samples belonging to the designated class label for a layer are assigned as positives while the remaining samples are negatives. The final auxiliary node for the layer is trained until all the samples have been correctly learned.

During detection time the prediction for a given sample is reached by evaluating it against individual nodes which return a vector of all possible class labels and their associated confidences. The confidences represent the sum of all class confidence labels which registered a hit from each weak classifier. The class label with the highest confidence sum is selected as the winning label.

The runtime classification of a given sample is efficient compared to most other approaches. While ECOC-based, OAO and OAA approaches to multiclass training require that all classifiers be executed in order to formulate a prediction, our classifiers undergo the calculation of only a subset of their complete makeup. As an unseen sample enters the cascaded structure, it will be ejected from a layer if any member node classifies it as matching its label, after which time the sample will be forwarded to the next layer until it reaches the layer that matches its class label. In this case, if the sample is correctly classified, it will be evaluated by every multiclass node within a layer. All multiclass nodes will reject the given sample as a negative while only the final auxiliary node will accept it as matching the layer class label. Additionally, as a sample instance propagates deeper into the cascade, the classification accelerates since there are a decreasing number of nodes at each step.

3 Experiments

We evaluated our algorithm on seven benchmark multiclass datasets from the UCI machine learning repository. We implemented AdaBoost.OC and ECC to compare with the proposed algorithm, while making use of existing results from [11] for AdaBoost.M2. For datasets with both training and test sets, we ran the experiments ten times; otherwise, 10-fold cross-validation was employed in conjunction with 10 training repetitions for a total of 100 runs. All results were averaged and in the presence of randomness, standard error was reported. For the multiclass cascade, we trained four different classifiers for each dataset with different parameter settings for determining the maximum

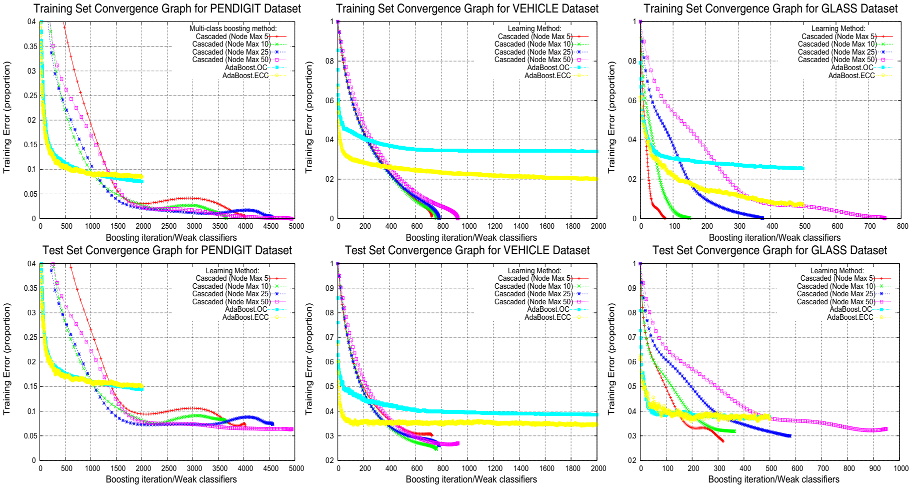


Fig. 2. Training and test error graphs for Pendigit, Vehicle and Glass datasets

number of weak classifiers per multiclass node. The sizes were 5, 10, 25 and 50. Each cascade layer target was set to a 100% hit rate for the designated class label and a 0% false alarm rate for every final auxiliary node. The terminating criterion for the entire cascade was zero training error. For a fair comparative analysis, AdaBoost ECC, OC and M2 classifiers were trained using decision stumps, while the total number of boosting iterations was in line with experiments in [11,7], which are reported in the results section.

We first examine the learning effectiveness and the generalization ability of our method before analyzing the runtime performances. In Figure 2, the training and test convergence patterns are illustrated for three selected datasets. It is clearly observable that during training, OC and ECC classifiers converge considerably faster in the initial boosting iterations; however, the cascaded classifiers catch up and eventually reach zero training error unlike their counterparts. This pattern was consistent across all datasets. Particularly on larger training datasets, the OC and ECC classifiers converged more rapidly to a given point from which the training error decreased marginally, and reached zero training error only once. Stagnation in the subsequent convergence indicated the inability of OC and ECC approaches to improve learning on challenging datasets when given a weak learner and a naive training architecture. On the other hand, while the training cost of the cascaded classifiers was higher at onset, they converged to a zero training error with only one exception. This demonstrated the efficacy of our strategy to employ a stepwise decomposition of a training multiclass problem into cascades under direction of the boosting process. The method strengthens a weak base learner to the extent that arbitrarily low training error rates become achievable.

The test error graphs, seen also in Figure 2, mirror those of their associated training convergence patterns. Correspondingly, the accuracy rates of the OC and ECC classifiers is preferable over the cascaded classifiers when considering the initial boosting rounds; nevertheless, our experiments illustrated that in many cases, subsequent rounds

Table 1. Results from seven UCI datasets, featuring comparisons of error proportions on test sets, total numbers of boosting rounds per classifier and the detection runtime in seconds per sample to the accuracy of ±10%

		Multi-Class Cascaded				**ECC**	**M2**	**OC**
DATASET		Node Sizes					Eibl[11]	
		5	10	25	50			
PENDIGIT	test error	**0.067**	**0.071**	**0.067**	**0.053**	0.15 ±0.01	0.186	0.144 ±0.002
	boosting iterations	(4157)	(3722)	(4715)	(5085)	(2000)	(2000)	(2000)
	execution runtime	**4.3e-05**	**6.1e-05**	1.2e-04	2.2e-04	8.5e-04	-	1.0e-04
SATIMAGE	test error	**0.152**	**0.145**	**0.139**	**0.147**	0.228 ±0.004	0.182	0.163 ±0.002
	boosting iterations	(2621)	(2608)	(2760)	(2745)	(2000)	(2000)	(2000)
	execution runtime	**1.9e-05**	**2.6e-05**	**4.1e-05**	**6.9e-05**	4.0e-04	-	7.7e-05
VOWEL	test error	0.616	**0.541**	0.735	0.688	0.611 ±0.013	0.543	0.626 ±0.01
	boosting iterations	(1109)	(1368)	(2085)	(3419)	(2000)	(2000)	(2000)
	execution runtime	**4.3e-05**	**7.0e-05**	1.4e-04	2.8e-04	9.9e-04	-	1.1e-04
SEGMENTATION*	test error	0.0554 ±0.008	**0.0390 ±0.005**	0.0511 ±0.007	0.0459 ±0.004	0.081 ±0.014	0.084	**0.0468 ±0.017**
	boosting iterations	(814)	(978)	(1297)	(1513)	(2000)	(2000)	(2000)
	execution runtime	**1.3e-05**	**2.6e-05**	**5.2e-05**	**7.3e-05**	5.2e-04	-	9.2e-05
VEHICLE*	test error	0.299 ±0.016	**0.245 ±0.019**	**0.26 ±0.009**	**0.27 ±0.011**	0.345 ±0.046	0.353	0.386 ±0.03
	boosting iterations	(721)	(787)	(809)	(976)	(2000)	(2000)	(2000)
	execution runtime	**9.0e-06**	**1.3e-05**	**1.8e-05**	**2.8e-05**	2.7e-04	-	6.8e-05
GLASS*	test error	0.278 ±0.025	0.319 ±0.027	0.3 ±0.027	0.328 ±0.01	0.37 ±0.075	**0.25**	0.36 ±0.078
	boosting iterations	(309)	(360)	(592)	(985)	(500)	(500)	(500)
	execution runtime	**9.7e-06**	**1.6e-05**	3.5e-05	6.0e-05	1.1e-04	-	2.1e-05
IRIS*	test error	0.0733 ±0.018	0.0800 ±0.016	0.0867 ±0.019	**0.0533 ±0.018**	**0.079 ±0.037**	0.055	**0.066 ±0.027**
	boosting iterations	(50)	(65)	(123)	(228)	(500)	(500)	(500)
	execution runtime	**2.2e-06**	**3.8e-06**	**7.0e-06**	1.4e-05	4.8e-05	-	1.7e-05

contribute to an improved accuracy of the proposed algorithm over the OC and ECC methods, while being comparable in the rest. This is supported by the figures in Table 1, which lists the final accuracy rates of all classifiers across the seven datasets, with the best performing and statistically significant results highlighted.

Table 1 also reports the execution runtimes in seconds for all classifiers except M2. The best performing results are highlighted. The results show that all the fastest detection times are achieved by the cascaded classifiers and in particular by those with the smallest node sizes. It is significant, that in most cases the most accurate classifiers have also registered faster runtime performances. Out of ECC and OC classifiers, consistently faster runtimes were archived by OC classifiers. The accelerated execution runtime of the cascaded classifiers can be explained by the fact that only a subset of the entire ensemble requires evaluation per instance, in contrast to the naive implementation of monolithic ensembles of ECC and OC. Since each layer in a cascade is embedded with multiple exit points, most layers will only be exposed to partial execution even if the target class label for a particular candidate instance is assigned to the last layer. Moreover, as a candidate sample propagates through a cascade, the size of the layers also decrease in size; thus, further minimizing the execution time.

We observed that further execution runtime increases can be gained. Currently the performance bottleneck lies with the evaluation of the domain partitioning of the multiclass weak classifiers which increases linearly with the number of classes. This can be seen from Glass and Iris datasets where the cascaded classifiers comprised of significantly smaller ensembles, yet only performed marginally better than ECC and OC classifiers. In these instances, the ratio of multiclass weak classifiers greatly outweighed the binary weak classifiers. We believe that by implementing lookup tables at detection time for each multiclass node, a significant improvement can be realized.

4 Conclusion

In this paper we have presented a unique cascaded training algorithm for multiclass ensemble-based problems, as well as a new weak learning method to accompany it. We demonstrated the ability of our strategy to create robust and real-time capable classifiers that are appropriate for time-critical computer vision applications with faster runtimes than AdaBoost.OC and ECC.

Most multiclass algorithms based on weak learners cannot achieve arbitrarily low training errors on difficult datasets and therefore resort to computationally more complex learners, which results in protracted training runtimes. Using seven benchmark UCI datasets, we have show how a simple weak learner can be combined with a multiclass decomposition strategy that organizes an ensemble into cascades. We demonstrate the ability of our approach to reach arbitrarily low training errors while displaying stronger generalization rates in five out of seven datasets compared to AdaBoost.OC, ECC and M2 when using very weak learners.

References

1. Lorena, A.C., Carvalho, A.C., Gama, J.a.M.: A review on the combination of binary classifiers in multiclass problems. Artif. Intell. Rev. 30, 19–37 (2008)
2. Verschae, R., del Solar, J.R.: Coarse-to-fine multiclass nested cascades for object detection. In: International Conference on Pattern Recognition, pp. 344–347 (2010)
3. Li, L.: Multiclass boosting with repartitioning. In: Proc. of the 23rd Int. Con. on Machine Learning, ICML 2006, pp. 569–576. ACM, NY (2006)
4. Allwein, E.L., Schapire, R.E., Singer, Y.: Reducing multiclass to binary: a unifying approach for margin classifiers. J. Mach. Learn. Res. 1, 113–141 (2001)
5. Dietterich, T.G., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. CoRR cs.AI/9501101 (1995)
6. Guruswami, V., Sahai, A.: Multiclass learning, boosting, and error-correcting codes. In: Proc. of the 12th Ann. Conf. on Comput. Learn. Theory, COLT 1999, pp. 145–155. ACM, New York (1999)
7. Freund, Y., Schapire, R.: Experiments with a new boosting algorithm. In: Proceedings of the 13th International Conference in Machine Learning, pp. 148–156 (1996)
8. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1), 119–139 (1997)
9. Viola, P., Jones, M.: Robust real-time face detection. In: Proc. of the 8th Int. Con. on Computer Vision (ICCV 2001), p. 747. IEEE Computer Society, Los Alamitos (2001)
10. Schapire, R.E., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. Machine Learning 37, 297–336 (1999)
11. Eibl, G., Pfeiffer, K.P.: Multiclass boosting for weak classifiers. J. Mach. Learn. Res. 6, 189–210 (2005)

Mutual Information Based Gesture Recognition

Peter Harding, Michael Topsom, and Nicholas Costen

School of Computing, Mathematics and Digital Technology,
Manchester Metropolitan University, UK
{p.harding,m.topsom,n.costen}@mmu.ac.uk

Abstract. Proliferation of gestural interfaces necessitates the creation of robust gesture recognition systems. A novel technique using Mutual Information to classify gestures in a recognition system is presented. As this technique is based on well-known information theory metrics the underlying operation is not as complex as many other techniques which allows for this technique to be easily implemented. A high recognition rate of 98.55% was achieved, with recognition occurring in under 10ms.

Keywords: Mutual Information, Pattern Recognition, Classification, User Interface.

1 Introduction

The recent proliferation of touch screen, accelerometer based, haptic, and other gestural interfaces necessitates the creation of robust gesture recognition systems to ensure their fast and reliable operation. The inclusion of these interfaces in modern electronic devices (e.g. mobile phones, hand-held touch devices [15], and computer game consoles [2]), which often have access to limited processing power, requires these recognition systems to be computationally efficient to allow the classification of input at a near real-time speed which is considered acceptable to users [10]. This paper presents a lightweight, simple to implement recognition system, based on information theory techniques, which fulfils these criteria, and details the results of testing as to illustrate the effectiveness of this system.

2 Recognition Problem

The recognition problem addressed is that of the correct classification of two-dimensional glyphs [12], of the type routinely used as control input for touch screen, stylus or wand driven devices ([8] gives an example control interface). A set of sixteen gestural input glyphs is employed, as seen in Figure 1, which has previously been used (in whole or in part) during the testing of this type of system [7,16], and is believed to offer a reasonable cross section of the possible gestures that would be found in modern user interfaces.

The recognition of these glyphs must be shown to be robust, as even a single user system may have to deal with "noisy" input, for various reasons [17]. The recognition provided must also be shown to be computationally efficient, allowing recognition at a rate that may be considered to be in real time, from the user's perspective.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 571–578, 2011.

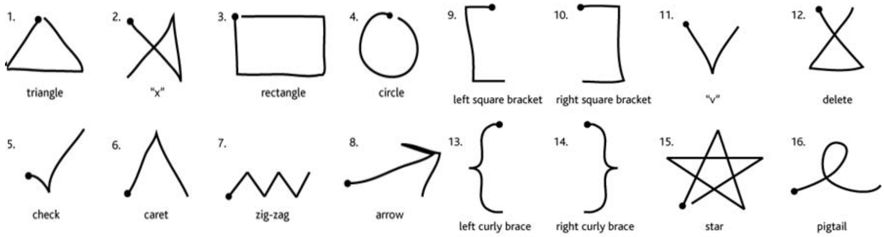


Fig. 1. Unistroke gestures

3 Recognition Using Mutual Information

The proposed technique has two basic sections, the first of these concerns the processing of raw data to transform it to a more usable form. The data is then passed to the classification system, which performs the comparisons against a template bank.

The data capture system used is touch screen based, and reads user input as a set of N Cartesian coordinates sampled from a single continuous motion. This method was chosen as it is analogous to myriad accelerometer, wand and mouse based interfaces to modern electronic devices.

3.1 Pre-processing

The initial processing steps performed are not uncommon in classification systems, and consist of the re-sampling, rotation and scaling of the raw data.

User input was found to be of varying size (i.e. the number of points), due to factors such as the speed with which the gesture was made and the data capture technique. The first pre-processing step normalizes the number of points. The raw data are re-sampled, interpolating to ensure that all points are of a fixed size and equidistantly spaced, leaving a vector of points, N'. Figure 3.1 shows a re-sampling of a single input of size N to create a processed set of points at size of N'. The second pre-processing stage rotates the gesture based on the angle between the first recorded point of the input, and the centroid of the input, see Figure 3.1, so the angle is uniformly 0^c. This mitigates any error caused by poorly orientated input, and is required to ensure the robustness of the recognition technique. Finally the gestures are scaled to have a fixed bounding box. Each (x, y) value is transformed to lie within the range $\pm\kappa$ (an arbitrarily chosen scaling constant), where

$$v' = 2\kappa \left(\frac{v - min(V)}{max(V)} \right) - \kappa, \quad v \in V. \tag{1}$$

This is applied separately to the (x, y) values (v is an arbitrary symbol) and $\kappa = 1$.

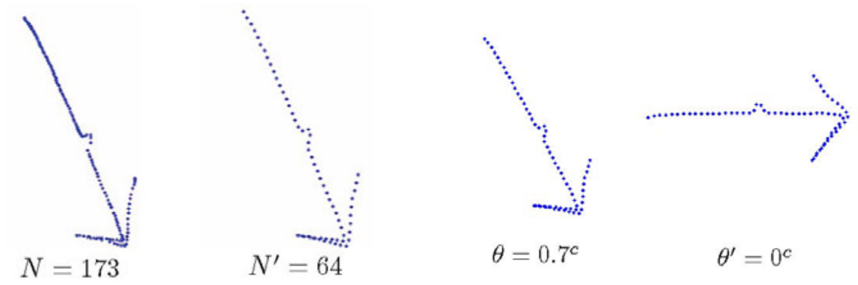


Fig. 2. Resampling and subsequent rotation of an input glyph

3.2 Mutual Information Analysis

The actual recognition method for these glyphs is based on Mutual Information (MI), a probabilistic method for quantifying the interdependence of two signals. It has previously been employed as an analytical technique in many areas [3,5,4], including classification tasks [1,14], but appears not to have been applied to the problem of gesture recognition.

The weighted mutual information of two *discrete* time-series variables, T and U, is defined as

$$I(T;U) = \sum_{i,j} w(t_i,u_j)P(t_i,u_j)\log_n \frac{P(t_i,u_j)}{P(t_i)P(u_j)} \tag{2}$$

where $P(t_i)$, $P(u_j)$, and $P(t_i,u_j)$ are the individual and joint probability distributions of T and U respectively. In general terms, the MI of two signals quantifies their interdependence; therefore if T and U are entirely independent, then $I(T;U) = 0$, but in *all* other cases $I(T;U) > 0$. The use of weights $w(t_i,u_j)$ in MI can increase recognition accuracy [11,6]. This scales $I(t_i,u_j)$ either upwards if $t_i \approx u_j$, or downwards if $t_i \neq u_j$. This creates a reward structure for correct values, whilst penalising any pairs of values that are not correctly identified.

The weighting function employed in this paper is a Gaussian distributed function of the absolute difference of the two input values, scaled by σ. Initial experimentation showed that the application of the weighting matrix improved results considerably, but only for very small variances, so $\sigma^2 = 10^{-2}$, while $\kappa = 1$.

User input is read in the form of a vector of Cartesian coordinates, U, which is then re-sampled, rotated and scaled as described previously. These coordinates are then separated into their x and y components, and discretised into R equally sized bins $R \in \{3,\ldots,9\}$, leaving two discrete vectors U_x and U_y. The each set of template data, T, is processed in exactly the same manner. The mutual information, I, is calculated as $I = I(T_x;U_x) \times I(T_y;U_y)$.

4 Experimentation and Results

Experimentation was carried out in three stages; ideal data recognition, user data (including comparison with an existing system) and additional noise. These testing stages were designed to test the limits of the system under different circumstances. Idealised data testing shows performance with known inputs and parameterized variations. The user data testing tests the ability of the method to classify gestures in a real world context. The addition of noise to data tests the ability of the system to recognise and correctly classify distorted data, which is an important test for any system that may not be deployed in an ideal scenario.

4.1 Ideal Data Recognition

A set of *perfect patterns* (precisely defined, uniform inputs) were created, which consisted of five points joined by four straight lines. The position of the final point of the pattern was then repositioned to a total 121 different locations, which were uniformly distributed inside the pattern, to create a test set. An example of one of these patterns may be seen in Figure 3.

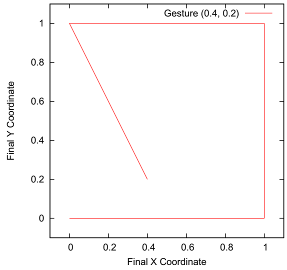


Fig. 3. The "ideal data" pattern; the outer lines enforce orientation. The variable point is at (0.4,0.2).

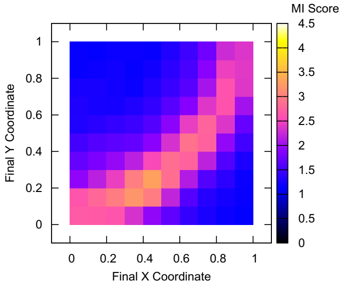


Fig. 4. Variation in recognition performance with location of the point. The probe point is at (0.4,0.2).

To ensure that the system could recognise data reliably *all* of the 121 data items were both used as a test set and a template set for these experiments. The system was presented with each of the 121 data items in turn, and then logged both the classification given and the MI score returned at $R = 7$. The system achieved a 100% accuracy in classification during these tests, i.e. each input presented was identified as its corresponding data item from the template bank. This shows that the system is able to accurately distinguish between large sets of relatively similar gestures and retain a good degree of accuracy.

To investigate the variance in MI results across increasingly distorted versions of the same input, the MI scores returned when the input shown in Figure 3 was compared with *all* of the 121 templates. Figure 4 shows that higher recognition values lie on the arc where the pattern retains exactly the same length, i.e. the points at which the length of the fourth line in the pattern retains the same length as in the recognition template. When two ideal data patterns are of the same length re-sampling will produce many corresponding points along the first three straight lines increasing the MI score.

4.2 User Data

Sets of test data, consisting of three examples of each of the sixteen Figure 1 glyphs, were collected from 26 test subjects. Four data items were not recorded due to experimenter error, resulting in a total of 1244 unique data items being used for testing. The testing was carried out using a *leave-one-user-out* testing strategy; in turn, each user was supplied probes, and all remaining 25 sets of user data formed the gallery.

Considering the fast and lightweight nature of the MI system, a suitable comparison is the $1 recogniser [16] (which uses a geometric measure for classification and utilises the same pre-processing steps). This has been shown to operate at a faster speed than both a Rubine Classifier [13] and a Dynamic Time Warping based matcher [9], and has at worst comparable but often more accurate recognition to these systems. The same gesture set and testing strategy were employed. The recognition rate and recognition speed was recorded for various re-sampling values of N with both systems.

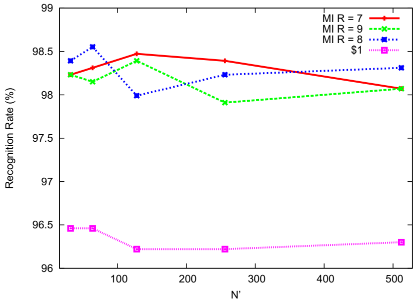


Fig. 5. Comparison of the accuracy of the MI recognition system ($7 \leq R \leq 9$) and $1 recogniser

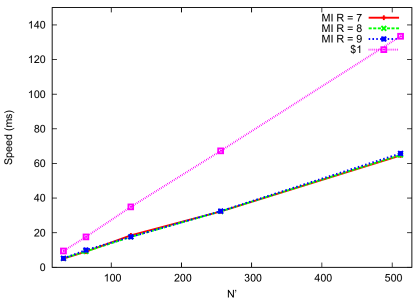


Fig. 6. Comparison of the speed of the MI recognition system ($7 \leq R \leq 9$) and $1 recogniser

The accuracy of the MI technique reached a maximum recognition rate of 98.55% while the $1 recogniser reached a maximum recognition rate of 96.46%. The confusion matrix in Figure 7 shows the results for the MI classifier. For all values of $R > 3$ the recognition technique out performed the $1 recogniser in classification and speed. Figure 5 shows the MI technique's highest recognition rates ($7 \leq R \leq 9$) compared to the absolute highest recognition rate achieved by the $1 recogniser.

The speed at which each technique recognised and classified an input gesture was recorded, this was calculated by sequentially recognising each of the 1244 data items and averaging the net time taken. Experiments were run on a desktop computer with Intel® Core^{TM}2 Quad CPU Q2800 running at 2.33GHz and 4GB of RAM with Java version 1.6.0_11. Figure 6 shows the speeds at which the MI system performed recognitions (binning values $7 \leq R \leq 9$; note these render to the same line). The MI system took approximately half the time to perform a classification that was required by the \$1 recogniser for a give value of N'. The value of R was found to have little effect on the speed of the MI system in comparison with the value of N'.

4.3 Additional Noise

To further investigate the robustness of the recognition technique, a series of experiments were run in which with additional noise applied to the probes before classification. The noise, η, was applied according to a directed, Gaussian distributed function

$$\eta_{n+1} = \eta_n + d\,|(\mathcal{N} : \mu, \sigma)| \quad d \in \{-1, 1\} \tag{3}$$

where d defines the directionality of the noise, and will change with a probability of $P(d_{n+1} = -d_n) = \frac{N}{2}$, yielding one expected change in the directionality of the noise for each probe. Noise is applied independently to both the x and y components of the signal, so at any time each component will have a separate and independent η. The noise is cumulative, which ensures that the signal will not be raised and lowered repeatedly; rather it will increased or decreased in a natural manner over time. This is arguably similar to the atypicalities found in human movements.

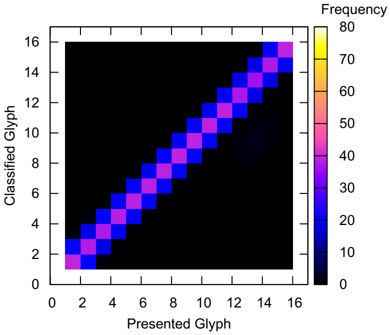


Fig. 7. Confusion matrix showing the probe glyph against the system's classification

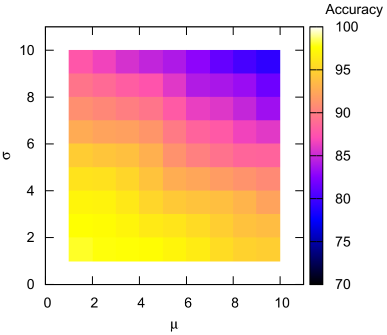


Fig. 8. Recognition accuracy of the MI system with varying values of μ and σ in the added noise

The same testing technique was employed across the new data set. The results of these experiments can be seen in Figure 8. The best recognition rates were achieved at low σ and μ values, where the recognition rate peaked at the 98.55% recorded during the user testing experiments, and recognition rates show a steady decrease as both μ and σ is increased. Even at the largest values $\mu = 10$ and $\sigma^2 = 10$ (note: maximum bounds of the glyphs were approximately 350 by 350 pixels before processing) the lowest recognition rate recorded was still 75.8%, which is twelve times greater than the naïve rate for this template set.

5 Discussion and Conclusions

Mutual information has been shown to work well when applied to the recognition of 2D gestures. In this series of experiments the MI system was shown to classify gestures with a high degree of accuracy; with a 100% recognition rate on artificial gesture data and 98.55% with user gesture data. The addition of noise to the user data lowered the accuracy of recognitions, although a recognition rate of over 75% was achieved in the worst conditions.

The recognition system managed to perform recognitions, on average, in under 70ms in the worst cases (highest R and N' values). The optimum recognition rates (98.55%) were achieved in under 10ms. The limiting factor in terms of speed of recognition was found to be the re-sampling rate N', this is not considered a problem as the optimum value for N' will, in most circumstances, be dictated by the sampling rate of the hardware in question. In the case of the machine used during the testing covered in this paper the maximum number of samples collected for any gesture was less than 500 points and was regularly found to be lower than 200 points. It is safe to assume that systems that have a higher sampling rate are also likely to have more processing power available for the MI recognition technique itself. For a user interface to be seen as responsive by users, it is suggested that the system should respond in under 100ms [10], the MI based system fulfils this requirement amply. It is also significantly more accurate and faster than other algorithms.

6 Further Work

As the x and y components of the signals are analysed seperately this method can be simply extended to a third dimension, allowing for input from a 3-axis accelerometer based device. As the technique has been found to be so fast, a large template bank was used during these experiments; template reduction techniques may be adapted to further increase recognition speed, which could allow a whole new area of low computational power micro-devices to incorporate gesture based control techniques into their software.

Acknowledgements. The authors gratefully acknowledge support from the Dalton Research Institute and the EPSRC-funded Bridging the Gaps: Nano-Info-Bio project, grant reference EP/H000291/1

References

1. Bahl, L., Brown, P., De Souza, P., Mercer, R.: Maximum mutual information estimation of hidden Markov model parameters for speech recognition. Acoustics, Speech, and Signal Processing 11, 49–52 (1986)
2. Cummings, A.H.: The evolution of game controllers and control schemes and their effect on their games. In: The 17th Annual University of Southampton Multimedia Systems Conference (2007)
3. Fraser, A.M., Swinney, H.L.: Independent coordinates for strange attractors from mutual information. Physics Review A 33(2), 1134–1140 (1986)
4. Harding, P.J., Amos, M., Gwynne, S.: Mutual information for the detection of crush. In: Proc. 4th Intl. Conference on Pedestrian and Evacuation Dynamics (2010)
5. Jeong, J., Gore, J.C., Peterson, B.S.: et al. Mutual information analysis of the EEG in patients with Alzheimer's disease. Clinical Neurophysiology 112(5), 827–835 (2001)
6. Junli, L., Rijuan, C., Linpeng, J., Ping, W.: A medical image registration method based on weighted mutual information. Bioinformatics and Biomedical Engineering, 2549–2552 (2008)
7. Chris Long Jr., A., Landay, J.A., Rowe, L.A., Michiels, J.: Visual similarity of pen gestures. In: CHI 2000, pp. 360–367 (2000)
8. Moyle, M., Cockburn, A.: The design and evaluation of a flick gesture for 'back' and 'forward' in web browsers. In: Biddle, R., Thomas, B.H. (eds.) AUIC. CRPIT, vol. 18, pp. 39–46. Australian Computer Society (2003)
9. Myers, C.S., Rabiner, L.R.: A comparative study of several dynamic time-warping algorithms for connected word recognition. The Bell System Technical Journal 60(7), 1389–1409 (1981)
10. Nielsen, J.: Usability Engineering. Morgan Kaufmann Publishers Inc., San Francisco (1995)
11. Novovičová, J., Somol, P., Haindl, M., Pudil, P.: Conditional mutual information based feature selection for classification task. In: Rueda, L., Mery, D., Kittler, J. (eds.) CIARP 2007. LNCS, vol. 4756, pp. 417–426. Springer, Heidelberg (2007)
12. Rubin, J.: Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests. Wiley, Chichester (1994)
13. Rubine, D.: Specifying gestures by example. In: SIGGRAPH 1991, pp. 329–337 (1991)
14. Shan, C., Gong, S., McOwan, P.W.: Conditional mutual information based boosting for facial expression recognition. In: BMVC, vol. 1, pp. 399–408 (2005)
15. Wilson, A.D.: Touchlight: an imaging touch screen and display for gesture-based interaction. In: Sharma, R., Darrell, T., Harper, M.P., Lazzari, G., Turk, M. (eds.) ICMI, pp. 69–76. ACM, New York (2004)
16. Wobbrock, J.O., Wilson, A.D., Li, Y.: Gestures without libraries, toolkits or training: a $1 recognizer for user interface prototypes. In: Shen, C., Jacob, R.J.K., Balakrishnan, R. (eds.) UIST, pp. 159–168. ACM, New York (2007)
17. Yee, W.: Potential limitations of multi-touch gesture vocabulary: Differentiation, adoption, fatigue. In: Jacko, J.A. (ed.) HCI International 2009. LNCS, vol. 5611, pp. 291–300. Springer, Heidelberg (2009)

Logitboost Extension for Early Classification of Sequences

Tomoyuki Fujino[1], Katsuhiko Ishiguro[2], and Hiroshi Sawada[2]

[1] Graduate School of Science and Technology, Keio University
fujino@thx.appi.keio.ac.jp
[2] NTT Communication Science Laboratories
{ishiguro.katsuhiko,sawada.hiroshi}@lab.ntt.co.jp

Abstract. We propose a new boosting method for classification of time sequences. In the problem of on-line classification, it is essential to classify time sequences as quickly as possible in many practical cases. This type of classification is called "early classification." Recently, an Adaboost-based "Earlyboost" has been proposed, which is known for its efficiency. In this paper, we propose a Logitboost-based early classification for further improvements of Earlyboost. We demonstrate the structure of the proposed method, and experimentally verify its performance.

Keywords: Time Sequence classification, Logitboost, Early Recognition.

1 Introduction

Classification of time sequence data is one of the most important problems in machine learning and computer vision, and is applicable to many practical problems including on-line handwriting recognition [1] and the classification of human behaviors [12]. Time sequence classification is typically tackled using generative models such as HMM [11]. These models require the entire sequence data $\{x(\tau)\}_{\tau=1}^{T}$ in general, but making a classification decision at an early time t on a time sequence $1 \leq t < T$ is desirable in many practical applications. On the other hand, many classification methods that are based on the discriminative models have been intensively developed in recent years. Support Vector Machine [14], Adaboost [2] and Conditional Random Field [8] are well known for their high performance, and have been employed in several studies [7,5].

Our problem is determining the class of a sequence of length T at an early time $t \leq T$ by using only the early time dataset $\{x(\tau)\}_{\tau=1}^{t}$; this problem is called "early classification" in this paper. In other words, the early classification problem is to classify a time sequence as quickly as possible, before the sequence ends. For example, consider the problem of driver behavior recognition from images captured by a camera installed in a vehicle [12] for driver assistance systems that make driving comfortable and safe. If we detect a sign of dangerous movements such as mobile phone use while driving, we would like to warn the driver quickly before the behavior finishes and causes any accidents.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 579–588, 2011.

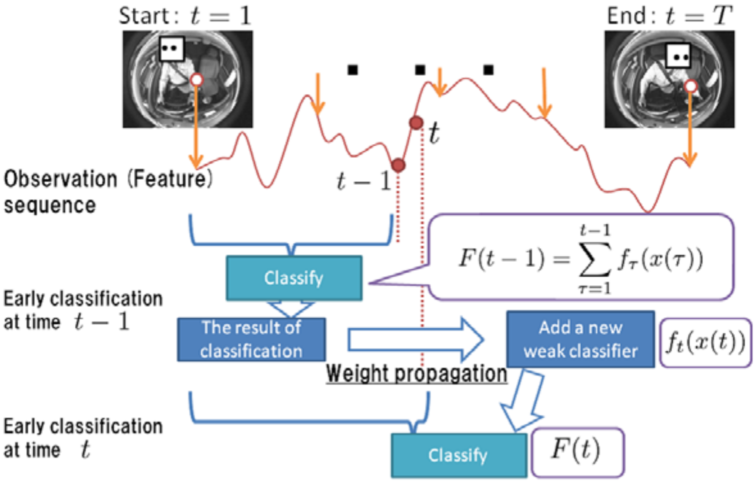

Start: $t = 1$
End: $t = T$
Observation (Feature) sequence
$t-1$
t
Classify
$F(t-1) = \sum_{\tau=1}^{t-1} f_\tau(x(\tau))$
Early classification at time $t-1$
The result of classification
Weight propagation
Add a new weak classifier
$f_t(x(t))$
Early classification at time t
Classify
$F(t)$

Fig. 1. Outline of early classification by boosting. Boosting generates a powerful classifier at each time t, and updates by adding an optimized weak classifier.

Recently, an Adaboost-based method called "Earlyboost" [13] has been proposed for early classification. In [6], the multi-class version of the method is also discussed with statistical background, and promising results are reported. However, these models are based on the original Adaboost [2]. Though the Adaboost is known as a powerful discriminative model, there must be many ways to improve the performance of the system, as the Adaboost has been modified by many researchers.

In this paper, we propose an extension of Earlyboost by using Logitboost [3]. Logitboost is an extension of Adaboost that improves classification performance by the maximum likelihood approach with logistic regression. Since the Earlyboost is based on the Adaboost, we expect further improvements of early classification by adopting the Logitboost algorithm (Fig. 1).

In section 2, we briefly review the previous boosting methods. In section 3, we propose a Logitboost-based early classification technique for the binary classification problems and the multi-class classification problems. The fourth section is devoted to experiments, and section 5 concludes this paper.

2 Background

2.1 Adaboost and Earlyboost

First, we briefly review Adaboost. Let $x_i \in \mathbb{R}^d$ and $y_i \in \{1, -1\}$ denote the training data sample and its corresponding label, respectively. N is the number of sample-label pairs. Adaboost constructs committee function $F(x)$ by combining a number of weak classifiers $f_m(x) : \mathbb{R}^d \rightarrow \{-1, 1\}$ as

$$F_M(x) = \sum_{m=1}^{M} c_m f_m(x), \tag{1}$$

where c is called importance weight and m indexes the weak classifier.

The key elements of Adaboost are the propagation of weight variables. Denoting w_i as the propagation weight for i-th datum, Adaboost increases the weights w_i of samples that are misclassified by the weak classifier at each iteration m (i.e. $y_i \neq f_m(x_i)$) by $w_i \leftarrow w_i \exp(c_m)$.

Theoretically, Adaboost is derived by minimizing the "exponential loss function" [3], which is defined as follows:

$$J(F) = \frac{1}{N} \sum_{i=1}^{N} \left[\exp(-y_i F(x_i))\right]. \tag{2}$$

Given $l-1$ iterations have finished, Adaboost optimizes the next weak classifier $f_l(x)$ and importance weight c_l by using the previous step's weight w_i. The minimization of (2) is boiled down to the following equation by a Taylor expansion:

$$f_l(x) = \arg\min_{f(x)} \frac{1}{N} \sum_{i=1}^{N} [-w_i y_i f(x_i)]. \tag{3}$$

The pseudo code of Adaboost is shown in Algorithm (1).

Algorithm 1. Standard Adaboost

```
Initialize $w_i = \frac{1}{N}, i = 1, \dots, N$.
for $m = 1, 2, \dots, M$ do
    (1)Fit the weak classifier $f_m(x) \in \{-1, 1\}$ using weight $w_i$ on the training dataset.
    (2)$\epsilon_m = \mathrm{E}_w[1_{(y \neq f_m(x))}]$.
    (3)$c_m = \log\left(\frac{(1-\epsilon_m)}{\epsilon_m}\right)$.
    (4)$w_i \leftarrow w_i \exp[c_m 1_{y_i \neq f_m(x_i)}], i = 1, 2, \dots, N$.
    (5)$w_i \leftarrow \frac{w_i}{\sum_{i=1}^{N} w_i}$.
end for
Output sign$[F(x)]$, where $F(x) = \sum_{m=1}^{M} c_m f_m(x)$.
```

Earlyboost is an application of Adaboost for early classification [13], which is similar to Adaboost except that the input samples $\{x_i(t)\}_{i=1}^{N}$ are quite different in each time frame, indexed by $t = 1, 2, \dots, T$. The strong classifier F is defined as an additive model of frame-dependent weak classifiers f_t.

$$F_T(x_i) = \sum_{t=1}^{T} c_t f_t(x_i(t)), \tag{4}$$

where i indicates the index of a sequence sample.

Earlyboost considers the distribution of samples as independent in each time t. Hence the weak classifiers f_t are independent for each time frame t, each f_t classifies only the observations in time frame t $x(t) = \{x_i(t)\}_{i=1}^N$, and propagating the sample weight $w_i(t)$ connects time frames. Following the almost identical path of Adaboost, we can easily derive Earlyboost. The pseudo code of Earlyboost is presented at Algorithm (2). We can interpret this algorithm that each weak classifier f_t learns the classification boundary at time t to minimize the classification error induced by the information up to time $t-1$. Thus the resulting strong classifier will be good for early classification of sequences, even if the sequence is short ($\tau \leq T$).

Furthermore, Earlyboost.MH, which is applied to multi-class classification problems, has been proposed. Please see [6] for details of Earlyboost.MH.

Algorithm 2. Earlyboost

```
Initialize $w_i = \frac{1}{N}, i = 1, \ldots, N$.
for $t = 1, 2, \ldots, T$ do
    (1)Fit the weak classifier $f_t(x) \in \{-1, 1\}$ using weight $w_i$ on the training dataset.
    (2)$\epsilon_t = \mathrm{E}_w[1_{(y \neq f_t(x(t)))}]$.
    (3)$c_t = \log\left(\frac{(1-\epsilon_t)}{\epsilon_t}\right)$.
    (4)$w_i \leftarrow w_i \exp[c_t 1_{y_i \neq f_t(x_i(t))}], i = 1, 2, \ldots, N$.
    (5)$w_i \leftarrow \frac{w_i}{\sum_{i=1}^N w_i}$.
end for
Output the classifier sign$[F(x(1:\tau))]$ at any time $1 \leq \tau \leq T$
```

Incremental nature of Earlyboost may remind the reader of online boosting [10,4]. We briefly explain the difference between these techniques and Earlyboost (early classification problem). Online boosting models are truly on-line: these models update the entire classifier given a current input of an observation and its label (x_i, y_i). If the input data is time-stamped, the strong classifier will change adaptively to classify the current input $x_i(t)$ correctly. In Earlyboost, however, we only optimize f_t only at the t-th round of training, exploiting $x(t) = \{x_i(t)\}_{i=1,\ldots,N}$. And Earlyboost classifies the sequence data obtained so far, not the current input $x_i(t)$.

2.2 Logitboost

Logitboost is a boosting method based on logistic regression [3], and the weak classifiers $f(x) : \mathbb{R}^d \to \mathbb{R}$ are optimized by maximizing a likelihood. Since the weak classifier is optimized directly (without weight variables), the committee function is represented as $F(x) = \sum_{m=1}^M \frac{1}{2} f_m(x)$. Logitboost employs a "binomial likelihood function" defined by Eq. (5) instead of the "exponential loss" (2) in Adaboost, and the probability of $y_i = 1$ is estimated by Eq. (6).

Algorithm 3. Standard Logitboost

```
Initialize $F(x) = 0$ and probability estimates $p_i = \frac{1}{2}$, $i = 1, 2, \ldots, N$
for $m = 1, 2, \ldots, M$ do
    (1)Calculate the working response $z_i$ and weight $w_i$ by Eq. (8).
    (2)Fit the weak classifier $f_m(x)$ using weighted least squares as (7).
    (3)$F(x) \leftarrow F(x) + \frac{1}{2} f_m(x)$ and $p_i \leftarrow \frac{1}{1+\exp(-2F(x_i))}$.
end for
Output the classifier sign$[F(x)]$
```

$$L(F) = \frac{1}{N} \sum_{i=1}^{N} [-\log(1 + \exp(-2y_i F(x_i)))]. \tag{5}$$

$$p_i = \frac{1}{1 + \exp(-2F(x_i))}. \tag{6}$$

For maximizing the likelihood, we adopt Newton's method. Then, we obtain the following least square criterion to find the weak classifier $f_l(x)$:

$$f_l(x) = \arg\min_{f(x)} \sum_{i=1}^{N} w_i (z_i - f(x_i))^2, \tag{7}$$

$$z_i = \frac{y_i^* - p_i}{p_i(1 - p_i)}, \quad w_i = p_i(1 - p_i), \tag{8}$$

where $y_i^* = \frac{y_i + 1}{2}$. A standard Logitboost is summarized as Algorithm 3.

3 Logitboost Extensions for Early Classification

3.1 Binary Logitboost for Early Classification

We develop a sequential extension of Logitboost in a way similar to deriving Earlyboost from Adaboost in [6]. Assuming the distributions of input samples $x(t) = \{x_i(t)\}_{i=1}^{N}$ are independent in each time t, the framework of Logitboost is applicable for time sequences. In this case, t-th weak classifier f_t is only applied to $x(t)$, the samples on the same time frame. Using the Bayesian law, a posteriori probability of $y_i^* = 1$ at time t is given by

$$p(y_i^* = 1 | x(1:t)) = \frac{p(x_i(t) | y_i^* = 1) p(y_i^* = 1 | x(1:t-1))}{\sum_{k=\{0,1\}} p(x_i(t) | y_i^* = k) p(y_i^* = k | x(1:t-1))} \tag{9}$$

where $x(1:t) = \{x(\tau)\}_{\tau=1}^{t}$. Defining a sigmoid function $\sigma(a) = \frac{1}{1+\exp(-a)}$, the posteriori probability is represented as $p(y_i^* = 1 | x(1:t)) = \sigma(a(t))$, where

$$a(t) = \ln \frac{p(x_i(t) | y_i^* = 1)}{p(x_i(t) | y_i^* = 0)} + \ln \frac{p(y_i^* = 1 | x(1:t-1))}{p(y_i^* = 0 | x(1:t-1))}. \tag{10}$$

The first term of Eq. (10) is a function with respect to $x_i(t)$, which is denoted by $f_t(x_i(t))$, The second term with respect to the prior probabilities equals $a(t-1)$. Assuming the second term is zero at $t = 1$ $(a(0) = 0)$, we obtain the additive model as $a(t) = f_t(x(t)) + \sum_{s=1}^{t-1} f_s(x(s))$, which corresponds to the recursive Bayesian inference. Now, the binomial log-likelihood is represented as

$$L(f_{1:t}) = \frac{1}{N}\sum_{i=1}^{N}\left[-\log\left(1+\exp\left(-2y_i\left(\sum_{s=1}^{t-1} f_s(x_i(s)) + f_t(x_i(t))\right)\right)\right)\right]. \tag{11}$$

Assuming p_i as the prior at time t, the derivatives for Newton's method $F(x) \leftarrow F(x) - H^{-1}(x)s(x)$ are computed as follows:

$$s(x(1:t)) = \left.\frac{\partial L}{\partial f_t(x(t))}\right|_{f_t(x(t))=0} = \frac{2}{N}\sum_{i=1}^{N}(y_i^* - p_i), \tag{12}$$

$$H(x(1:t)) = \left.\frac{\partial^2 L}{\partial f_t(x(t))^2}\right|_{f_t(x(t))=0} = -\frac{4}{N}\sum_{i=1}^{N} p_i(1-p_i). \tag{13}$$

The weak classifier is determined by using the weighted least squares as (7). Finally, the Logitboost for early classification is summarized in Algorithm 4.

Algorithm 4. Binary Logitboost for early classification

```
Initialize $F(x) = 0$ and probability estimates $p_i = \frac{1}{2}$, $i = 1, \dots, N$.
for $t = 1, 2, \dots, T$ do
    (1)Calculate the working response $z_i$ and weight $w_i$ by Eq. (8).
    (2)Fit the weak classifier $f_t(x(t))$ using weighted least square as (7).
    (3)$F(x(1:t)) \leftarrow F(x(1:t-1)) + \frac{1}{2}f_t(x(t))$ and $p_i \leftarrow \frac{1}{1+\exp(-2F(x_i(1:t)))}$.
end for
Output the classifier sign$[F(x(1:\tau))]$ at any time $1 \le \tau \le T$
```

3.2 Multi-class Logitboost for Early Classification

Multi-class Logitboost [3] maximizes the multinomial likelihood given by

$$L(y^*, p) = \sum_{k=1}^{K} y_k^* \log p_k, \tag{14}$$

$$p_{k,i} = \frac{\exp F_k(x_i)}{\sum_{j=1}^{K}\exp F_j(x_i)} \tag{15}$$

where index $k \in \{1, \dots, K\}$ denotes class, K is the class cardinality, $y_k^* = \{0, 1\}$ is the label and F_k is the classifier of class k. The sequential multi-class Logitboost is derived in a similar way to deriving sequential binary Logitboost.

Choosing a base class $k = K$ arbitrarily, and considering the multi-class Bayesian law, the likelihood (14) is reformulated to the additive model as

$$L(g_{1:K,1:t}) = \frac{1}{N} \sum_{i=1}^{N} \left[\sum_{k=1}^{K-1} y^*_{k,i} \left(\sum_{s=1}^{t-1} g_{k,s}(x_i(s)) + g_{k,t}(x_i(t)) \right) - \log \left(1 + \sum_{k=1}^{K-1} \exp \left(\sum_{s=1}^{t-1} g_{k,s}(x_i(s)) + g_{k,t}(x_i(t)) \right) \right) \right], \tag{16}$$

where $g_{k,t}(x_i(t)) = \log p_{k,i} - \log p_{K,i}$; also note that $g_{K,t} = 0$ at all time t. For Newton updates, we differentiate (16) and obtain the following derivatives:

$$s_k(x(1:t)) = \frac{1}{N} \sum_{i=1}^{N} (y^*_{k,i} - p_{k,i}), \tag{17}$$

$$H_{j,k}(x(1:t)) = -\frac{1}{N} \sum_{i=1}^{N} p_{j,i}(\delta_{j,k} - p_{k,i}), \tag{18}$$

where $j, k = 1, \dots, J-1$. Approximating the Hessian matrix as diagonal, the updates are produced as

$$g_{k,t}(x_i(t)) \leftarrow \frac{y^*_{k,i} - p_{k,i}}{p_{k,i}(1 - p_{k,i})}. \tag{19}$$

Assuming the diagonal Hessian matrix, the model learns weak classifiers of K classes independently. Instead, we obtain computationally efficient solutions as Eq.(19). We can fit the weak classifiers by a weighted least squares regression of $g_{k,t}(x_i(t))$ with weight $w_i = p_{k,i}(1 - p_{k,i})$ without the base class K. However, the optimization of classifiers should not depend on choosing base class K, so the symmetrization of classifiers is necessary. The conclusive update procedure with symmetrization is given by

$$f_{k,t}(x_i(t)) = \frac{K-1}{K} \left(g_{k,t}(x_i(t)) - \frac{1}{K} \sum_{k=1}^{K} g_{k,t}(x_i(t)) \right) \tag{20}$$

where $g_{k,t},\ k = 1, \dots, K$ are computed by (19). The multi-class Logitboost for early classification is summarized in Algorithm 5.

4 Experiments

4.1 Settings

We conducted experiments with two kinds of tasks (Fig. 2). The first experiment was on-line handwriting recognition. Two datasets were taken from the "Kuchibue" database: one is the English alphabet and the other is Japanese "hiragana" characters. Each sequence corresponds to a trajectory of writing a

Algorithm 5. Multi-class Logitboost for early classification

```
Initialize $F_k(x) = 0$ and probability estimates $p_{k,i} = \frac{1}{K}$, $k = 1, \ldots, K$, $i = 1, \ldots, N$.
for t=1,2,..., T do
    (1)Calculate the working response $g_{k,t}$ by eq. (19) and weight $w_i = p_{k,i}(1 - p_{k,i})$.
    (2)Fit the weak classifier $f_{k,t}(x(t))$ into $g_{k,t}(x(t))$ using weighted least squares (7).
    (3)Set $f_{k,t} \leftarrow \frac{K-1}{K}(f_{k,t}(x(t)) - \frac{1}{K}\sum_{k=1}^{K} f_{k,t}(x(t)))$
    (4)Update $F_k(x(1:t)) = F_k(x(1:t-1)) + f_{k,t}(x(t))$ and $p_{k,i} \leftarrow \frac{\exp(F_k(x_i(1:t)))}{\sum_{j=1}^{K}\exp(F_j(x_i(1:t)))}$
end for
Output the classifier $\arg\max_k[F_k(x(1:\tau))]$ at any time $1 \leq \tau \leq T$
```

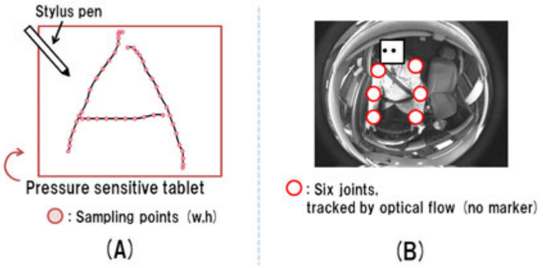

Stylus pen
Pressure sensitive tablet
○ : Sampling points (w,h)
(A)
○ : Six joints,
tracked by optical flow (no marker)
(B)

Fig. 2. The experiments. (A) Handwriting recognition task consists of English alphabet and Japanese hiragana recognition. (B) Driver behavior recognition task. Six joints are tracked by optical flow, without any markers.

complete character, and the length of the sequence was aligned to $T = 50$ by linear interpolations. The observed features $x_i(t)$ consisted of the 2D coordinates of a stylus pen tip on a pressure sensitive tablet and their velocity ($d = 4$). The performance of the methods was evaluated via 10-fold cross validation.

Our second experiment involved driving behavior recognition. A driver's behavior was recorded at 60 fps by a video camera installed in a driving simulator. Seven people participated in the experiment, and each person drove 30 times. The six joints of drivers were tracked by using optical flows, and the coordinates of joints (left and right wrists, elbows, and shoulders) were obtained on the 2D images. Thus, the feature $x_i(t)$ was $d = 24$ dimensional vector (12-dim. observations and 12-dim. velocities). The number of behavioral patterns was $K = 12$, which included "manipulating A/C," "adjusting the mirrors" and so on. All time frames were manually labeled, and segmented into behavior sequences. Each subject's behavior dataset is trained and evaluated separately by 6-fold cross validation, and the final result is averaged over seven subjects. The details of datasets are shown in Table 1.

4.2 Results

We compared the proposed multi-class sequential Logitboost with the previous method of Earlyboost.MH [6]. Both methods employed the "decision stump" as

Table 1. Details of datasets for each experiment

Task	alphabets	hiraganas	driving behavior
The number of classes K	52	83	12
The number of sequences N	14000	52500	$\simeq$ 650 for each subject
The dimension of features d	4	4	24

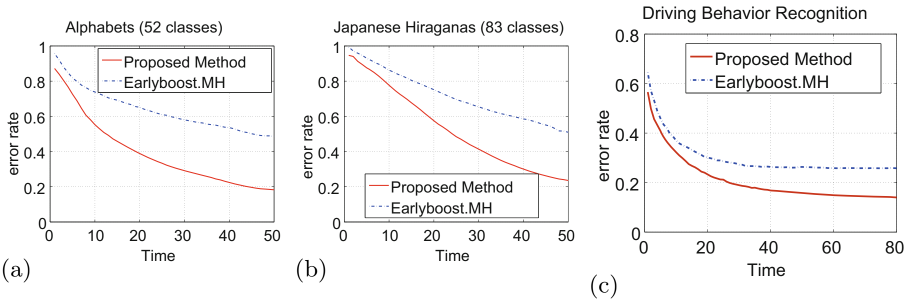


Fig. 3. Results of handwriting and driving behavior recognition. The red solid lines represent the averaged error rate of the early classification of the proposed method, and blue dashed lines represent those of Earlyboost.MH. (a) Alphabet (52 classes), (b) Japanese hiragana (83 classes), (c) Driving behavior recognition (12 classes).

a weak classifier. The averaged error rate of the test dataset of the alphabet task is shown in Fig. 3(a) where the vertical axis denotes the averaged error rate and the horizontal axis denotes the input sequences' time t. The hiragana recognition task's result is shown in Fig. 3(b), and the result of the driving behavior recognition task is shown in Fig. 3(c). The red solid lines denote the result by using the proposed sequential multi-class Logitboost, and the blue dashed lines denote the results by using Earlyboost.MH.

All results show that the proposed method is superior to Earlyboost.MH, which is based on Adaboost at all times. In general, Logitboost outperforms Adaboost due to the difference in optimization. In our early classification framework, however, we think the number of adaptable parameters also contributes to performance improvements. In Earlyboost.MH, the weak classifier is constrained to $f(x) : \mathbb{R}^d \rightarrow \{-1, 1\}$, and we have only T weight parameters $\{c_t\}_{t=1}^{T}$. On the other hand, the proposed Logitboost-based model optimizes the weak classifier $f(x) : \mathbb{R}^d \rightarrow \mathbb{R}$ directly: this corresponds to extending the Earlyboost.MH with $2 \times T \times K$ parameters. Therefore, Logitboost fit the weak classifiers into the maximizing likelihood better than Earlyboost.MH.

The objective of experiments is to measure the usefulness of Logitboost-based early classification against Earlyboost.MH, which is based on Adaboost. Thus we did not employ the state-of-the-art visual features such as SIFT [9], and utilized simple raw observations and velocities. We assume this is one reason that the classification rates remains not very high.

5 Conclusion

We presented a new boosting method that extends Logitboost for early classification of time sequences. We developed the formulations for both binary and multi-class problems by a Bayesian approach, and experimentally confirmed the superiority of the proposed method.

In this paper, we adopt the Logitboost for improving Adaboost-based early classification. However, many boosting models have been proposed by various researchers. Applying these models to the early classification will help more understanding of the problem.

References

1. Bahlmann, C., Burkhardt, H.: The writer independent online handwriting recognition system frog on hand and cluster generative statistical dynamic time warping. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(3), 299–310 (2004)
2. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computing Systems and Science 55(1), 119–139 (1997)
3. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: A statistical view of boosting (with discussion). Annals of Statistics 28(2), 337–407 (2000)
4. Grabner, H., Bischof, H.: On-line boosting and vision. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 260–267 (2006)
5. Gunawardana, A., Mahajan, M., Acero, A., Platt, J.C.: Hidden conditional random fields for phone classification. In: Proceedings of Interspeech (2005)
6. Ishiguro, K., Sawada, H., Sakano, H.: Multi-class boosting for early classification of sequences. In: Proceedings of the Twenty-first British Machine Vision Conference (BMVC), pp. 24.1–24.10 (2010)
7. Kim, K.J.: Financial time series forecasting using support vector machines. Neurocomputing 55, 307–319 (2003)
8. Lafferty, J., McCallum, A., Pereira, F.: Conditional random field: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of 18th International Conference on Machine Learning, pp. 282–289 (2001)
9. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
10. Oza, N., Russell, S.: Onlie bagging and boosting. In: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 105–112 (2001)
11. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
12. Sheikh, Y.A., Datta, A., Kanade, T.: On the sustained tracking of human motion. In: Proceedings of 8th IEEE International Conference on Automatic Face and Gesture Recognition, FG (2008)
13. Uchida, S., Amamoto, K.: Early recognition of sequential patterns by classifier combination. In: Proceedings of the 19th International Conference on Pattern Recognition, ICPR (2008)
14. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, Heidelberg (1995)

Determining the Cause of Negative Dissimilarity Eigenvalues

Weiping Xu, Richard C. Wilson, and Edwin R. Hancock

Dept. of Computer Science, University of York, UK
{elizaxu,wilson,erh}@cs.york.ac.uk

Abstract. Pairwise dissimilarity representations are frequently used as an alternative to feature vectors in pattern recognition. One of the problems encountered in the analysis of such data, is that the dissimilarities are rarely Euclidean, and are sometimes non-metric too. As a result the objects associated with the dissimilarities can not be embedded into a Euclidean space without distortion. One way of gauging the extent of this problem is to compute the total mass associated with the negative eigenvalues of the dissimilarity matrix. However,this test does not reveal the origins of non-Euclidean or non-metric artefacts in the data. The aim in this paper is to provide simple empirical tests that can be used to determine the origins of the negative dissimilarity eigenvalues. We consider three sources of the negative dissimilarity eigenvalues, namely a) that the data resides on a manifold (here for simplicity we consider a sphere), b) that the objects may be extended and c) that there is Gaussian error. We develop three measures based on the non-metricity and the negative spectrum to characterize the possible causes of non-Euclidean data. We then experimentally test our measures on various real-world dissimilarity datasets.

Keywords: non-Euclidean pairwise data, metric, embedding.

1 Introduction

Pairwise dissimilarity representations offer a powerful alternative to vectorial or feature-based characterisations of objects. Specifically, they provide a natural way of capturing the relationships between objects that are not characterised by ordinal measurements or feature vectors [6]. One way to translate such data into a vector representation is to represent the similarity data using a kernel matrix, and to embed the data into a vector space using kernel principal components analysis. In this way a vector representation is obtained by projecting the dissimilarity data into a vector space of fixed dimension.

However, one of the problems with dissimilarity representations and their embeddings is that the distance measures can not be used to construct a Euclidean vector space if the underlying Gram matrix contains negative eigenvalues. If this is the case, then the data can not be embedded into a real-valued Euclidean space, and must instead be embedded into a complex valued or Krein space [5].

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 589–597, 2011.

In order to analyse non-Euclidean dissimilarity data using traditional geometric machine learning or pattern recognition techniques, we must first attempt to rectify the data so as to minimize the non-Euclidean artifacts. Examples of translating similarities into vector representation include using only the positive definite subspace of the distances, adding a constant amount to the off diagonal elements, i.e. the constant shift embedding [4], or manifold embedding (e.g. the spherical embedding in [1]).

Each of these approaches is based on assumptions concerning the sources of the negative eigenvalues. The positive definite subspace embedding assumes that metric violations are an artifact of noise and that the distances in the negative sub-space do not carry any significant discriminative information. The manifold embedding assumes that the Euclidean violations are geodesic and that the data resides on a manifold. Recent studies [2,4] have showed that the negative eigenspace can contain valuable information. Moreover, Euclidean correction can lead to poor classification performance. Thus, before using any of the above approaches to attempt to rectify non-Euclidean data, it is advisable to analyze the underlying causes.

We model the distribution of non-Euclidean pairwise data in the following three situations: a) that the objects reside on the surface of a sphere (a simple manifold) and that the pairwise similarities are geodesic distances across the manifold, b) a non-metric dataset based on the distances between the surfaces of randomly positioned balls having different radii (Delft's balls data) and c) a noisy dataset with the Gaussian noise added to the distance between points in Euclidean space. By observing the spectrum of negative eigenvalues of the resulting Gram matrices and the additive constant required to render it metric, we identify three measures that can be used to characterise the above sources of negative eigenvalues. A variety of dissimilarity datasets are tested on the measures. Our analysis provides insight into the non-Euclidean behaviour of dissimilarity datasets and can be used to select appropriate embedding methods suited to the non-Euclidean data in hand.

Another secondary contribution of the paper is to develop a measure that assesses the contribution of each object to the mass of negative eigenvalues that provides further insight into the cause of non-Euclidean behavior. In this paper, we test a finer measure that assesses the contribution of each object to the mass of negative eigenvalues. In this way it is possible to determine whether the non-Euclidean artifacts are attributable to the dissimilarities associated with a few outlying objects or are uniformly distributed throughout the dataset.

2 Characterising the Causes of non-Euclidean Data

In this paper we are concerned with the sources of non-Euclidean data. Our overall aim is to identify the causes of a given set of non-Euclidean dissimilarity data so as to find out suitable correction methods to make them more Euclidean.

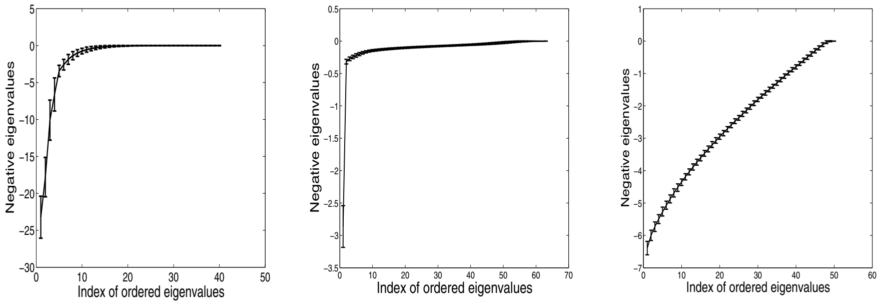
2.1 The Causes of non-Euclidean Data

We begin by identifying three reasons for non-Euclidean behaviour [2].

Manifold. If the data points reside on a curved manifold, then the distances between them are intrinsically non-Euclidean (but still metric). This is one possible source of non-Euclidean distances. Here we model such data as points on the surface of a sphere, a simple surface where distances are easy to compute. It is simple to simulate patches with various degrees of curvature that depart from Euclidean behavior by changing the curvature of the patch. The dissimilarity measurements on the sphere are metric but non-Euclidean.

Extended objects. If objects are not point-like but rather are extended in space, then the distances between them are measured between the closest points on their surface. As a result the distances will be non-Euclidean and possibly non-metric. Delft's balls data [2] is a typical example. Randomly positioned balls are generated with varying radius. The pairwise dissimilarities are the surface distances between the balls. As a result only the pairwise distances between balls with zero radius are Euclidean. It is also simple to modify the degree of non-Euclidean behaviour by adjusting the radii of the balls.

Gaussian noise. The final source is Gaussian noise added to the original Euclidean dissimilarities. This will generate data that is both non-Euclidean and non-metric.



(a) on the sphere (b) balls data (c) Gaussian noise

Fig. 1. The negative eigenvalues of the resulting Gram matrix of 100 points on the sphere, from extended objects and Gaussian noise as a function to the index of ordered negative eigenvalues

2.2 Negative Spectrum

We study with the three simple modes of the occurrence of non-Euclidean pairwise data. The Gram matrix of non-Euclidean dissimilarity data is indefinite, i.e. it has negative eigenvalues. One way to gauge the degree to which a pairwise distance matrix exhibits non-Euclidean artefacts is to analyse the properties of its centralised Gram matrix. For an $N \times N$ symmetric pairwise dissimilarity matrix D with the pairwise distance as elements, the centralized Gram matrix $G = -\frac{1}{2}JD^2J$,where $J = I - \frac{1}{N}11^T$ is the centering matrix and 1 is the all-ones vector of length N. The degree to which the distance matrix departs from being Euclidean can be measured by using the relative mass of negative eigenvalues or "negative eigenfraction " $F_{eigS} = \sum_{\lambda_i<0} |\lambda_i| / \sum_{i=1}^{N} |\lambda_i|$ [3]. This measure is zero when the distances are Euclidean and increases as the distance becomes increasingly non-Euclidean.

We commence by examining the negative spectrum of the Gram matrix under the three models. Figure 1 shows the non-Euclidean dissimilarities from the sphere and balls data-sets have spectrum which contain a strong negative component, with a concentration towards the low end of the spectrum. The non-Euclidean dissimilarities from Gaussian noise have a more slowly decreasing negative spectrum. Each of the negative spectrum appear to follow an exponential decay. Thus the slope and the intercept from an exponential fit should be able to discriminate at least the Guassian noise model from the remaining two models. An exponential curve of the form $y = ae^{bx}$ is fitted to the data, with b the slope and a the intercept. These two parameters are used as measures to characterise the negative spectrum.

2.3 Non-metricity

A distance measure is considered to be non-metric if it is either non-symmetric, negative or violates the triangle inequality. A dissimilarity matrix rarely satisfies the triangle inequality, but is usually positive [4]. Thus the violation of the triangle inequality is considered when measuring non-metricity. A constant $C = \max_{i,j,k} |d_{ij} + d_{ik} - d_{jk}|$ is computed and added to the off-diagonal elements of the dissimilarity matrix so as to increase the amount of data that satisfies triangle equality [3]. If C is zero, the pairwise dissimilarity is considered to be metric. Moreover, the dissimilarity values over the sphere are metric. Thus a, b and C can be used as three measures to identify the three modeled sources of non-Euclidean behavior.

2.4 Object's Contribution to the non-Euclidean Behaviour of Dissimilarities

If the non-Euclidean artefacts are created solely by the set of distances to a few outlying objects which are incorrectly placed, then it is possible to restore the data to a Euclidean state by editing these objects from the dataset. Based on this idea the notion of measuring the contribution of each object to the negative eigenfraction of a dissimilarity matrix is introduced.That is, the fraction given by the ratio of the sum of the negative distances originating from an individual object to each of the remaining objects, divided by the total.

The points can be embedded in Krein space as follows $Y = \sqrt{\Lambda}\Phi^T$ where Λ is the diagonal matrix with the ordered eigenvalues of centralised Gram matrix as elements and Φ is the eigenvector matrix with the ordered eigenvectors as columns. When the centered Gram matrix has negative eigenvalues then those dimensions of the embedding associated with negative eigenvalues are represented by imaginary numbers, and those associated with positive eigenvalues by real numbers. In other words, the data are embedded into a pseudo Euclidean or Krein space [5]. Under the embedding,the coordinate vector of point j is $y_j = (\sqrt{\lambda_1}\varPhi_{1j}, ..., \sqrt{\lambda_i}\varPhi_{ij}, \sqrt{\lambda_N}\varPhi_{Nj})^T$. The contribution to the squared distance between two points k and e is

$$d_{ke}^2 = \sum_i (y_k(i) - y_e(i))^2 = \sum_i \lambda_i (\phi_{ik} - \phi_{ie})^2$$

The sum of negative squared distances and the sum of positive squared distances from point k to all the remaining points are:

$$d_{k-}^2 = \sum_{\lambda_i<0} \lambda_i \sum_{e\neq k} (\phi_{ik} - \phi_{ie})^2, \quad d_{k+}^2 = \sum_{\lambda_i>0} \lambda_i \sum_{e\neq k} (\phi_{ik} - \phi_{ie})^2$$

Thus the fraction of negative squared distances from point k is

$$f_{pneig} = \frac{|d_{k-}^2|}{|d_{k-}^2| + |d_{k+}^2|}$$

This measure is zero for all objects (or points) when the distances are Euclidean and non-zero for outlier objects. Thus the measure can be useful to identify whether the non-Euclidean is caused by the second sources.

3 Experiments

To model distances sampled from a manifold, we commence with 100 points uniformly distributed on the surface of a 3D sphere with unit radius. The spherical coordinates of an object are $x = (r \sin\theta \cos\phi, r \sin\theta \sin\phi, r \cos\theta)^T$ where r is the radius of the sphere, θ is the elevation angle($[0, \pi]$) and $\phi([0, 2\pi])$ is azimuth angle. The pairwise geodesic distances are computed as the lengths of great circle arcs between pairs of objects. We can use the tangent space projection and increase the radius or change the range of the elevation angle to control the extent to which the patches deviate from a Euclidean surface, i.e. the degree of non-Euclideanness in the dissimilarity matrix. In total 100 initial configurations of points are used.

To model the extended objects, we pick 100 randomly positioned points in a 7D hypercube with length 100, and we take each point as the center of a ball with radius $r(r \geq 0)$. The balls do not overlap. The pairwise distance is the Euclidean distances between the centers of two balls minus the radii of the two balls. We regard the balls with radius greater than 0 as non-Euclidean balls. We vary the fraction of non-Euclidean balls, and take the fraction to be 0.1, 0,3, 0.5, 0.7 or 0.9 in our experiments. The radii of the non-Euclidean balls are 2, 3 or 4. We also generate 100 balls with uniformly distributed radii ranging from 0 to 4.

To model the Gaussian noise, we commence with 100 randomly positioned points in a 3D Euclidean space and calculate the Euclidean dissimilarity matrix. Then we add Gaussian noise with zero mean and various values of standard deviation to the off-diagonal elements of the dissimilarity matrix to generate a non-Euclidean dissimilarity matrix. The value of the standard deviation of the Gaussian noise is 0.1, 0.3, 0.5, 0.7 and 0.9.

To ensure the results are comparable over the dissimilarity data in various ranges and scales, all of the dissimilarity metrics are scaled such that the average dissimilarity is unity. We calculate the negative eigenvalues of each dissimilarity matrix and fit the average negative spectrum by an exponential curve to obtain the slope b, the intercept a and the average metric constant C. The whole process is repeated for a sample sizes of 500 and 1000 points.

Figure 2 shows the slope b as a function of the metric constant value C from the non-Euclidean dissimilarities on the sphere, the "balls" data and Gaussian noise. As the negative spectrum of the Gram matrix from the Euclidean points with Gaussian noise appears to be in a flat and linear in shape, so the value of slope b is very small with a value around -0.04. For the dissimilarities from the extended objects, the negative spectrum has a very sharp decreasing negative tail (just few significant negative eigenvalue), so the value for the slope b has a larger magnitude. Comparing the points on sphere and the ball data, there are several negative eigenvalues in the tail and the decrease is less sharp. This may explain why the slope of the non-Euclidean dissimilarities on the sphere is intermediate between that of the Gaussian noise and the non-Euclidean balls data. Another interesting finding is that the number of objects is not correlated with the slope, especially for points on the sphere and Gaussian noise. In terms of the parameters the three sources of negative eigenvalues are well separated from each other.

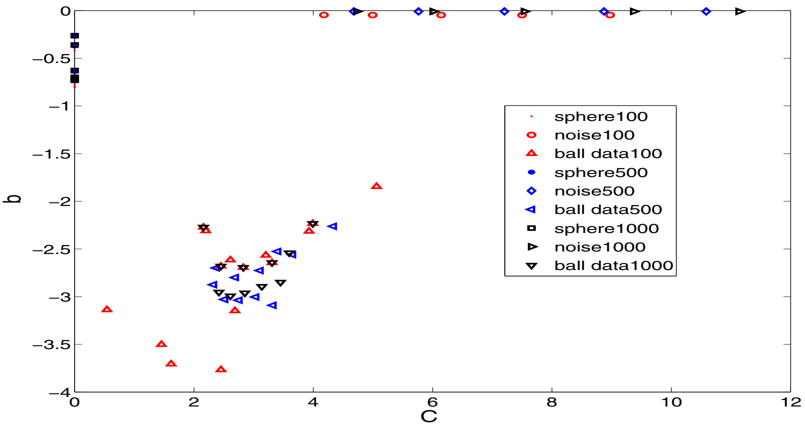
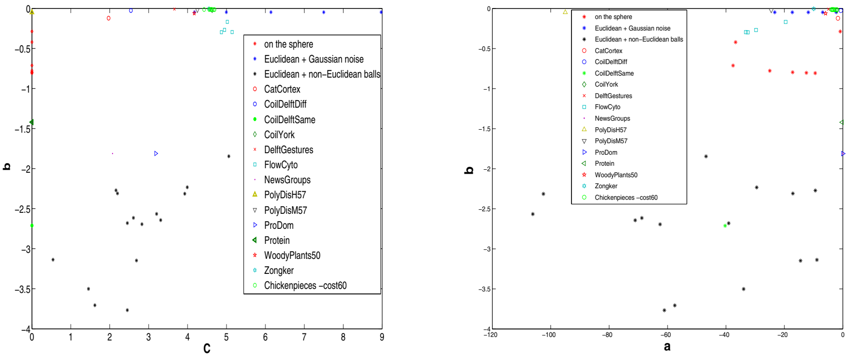


Fig. 2. The artificial non-Euclidean dissimilarity data caused by the manifold the data resides on, the extended objects and the Gaussian noise

We therefore use the above models to analyze a set of public domain dissimilarity data provided by the EU SIMBAD project consortium [2]. The Catcortex dataset contains dissimilarities based on the connection strengths between 65 cortical areas of the cat brain from four regions. CoilDelftDiff, CoilDelftSame and CoilYork are three dissimilariy datasets extracted from feature points detected in the COIL image database computed using different graph edit distances. FlowCyto contains four histogram dissimilarities for samples of breast cancer tissue. Newsgroups contains dissimilarities for messages in four classes of newsgroups. PolyDisH57 and PolyDisM57 are the dissimilarites of randomly generated polygons based on the standard and the modified Hausdorff distance. Protein contains the dissimilarities of protein sequences based on an evolutionary measure of distance. Woodyplants50 contains the shape dissimilarities between leaves of woody plants. Zongker contains the dissimilarities between handwritten digits based on deformable templates. Chickenpieces-cost60 contains 7 dissimilarity matrices from a weighed edit distance.



(a) (b)

Fig. 3. (a)The slope b as a function of the metric constant C; (b)The slope b as a function of the intercept a

The left and right plots in Figure 3 respectively show the slope b and the intercept a as a function of the metric constant value C, for the artificial samples of 100 objects. The plots indicate that the non-Euclidean behaviour of Defltgestures, PolyDisM57, Woodyplants50, Zongker, Chicken pieces, Catcortex and FlowCyto are likely to arise from Gaussian noise. On the other hand, the non-Euclidean behaviour of the Newgroups, ProDom and DelftSame datasets is likely to arise the non-Euclidean distances of a few outlying objects. We are unsure about the origin of the negative eigenvalues for the Protein and PolyDisH57 datasets. For PolyDisH57 the cause may be a combination of data residing on a manifold and the Gaussian noise. For the Protein dataset it may be a combination of data on the manifold and extended objects.

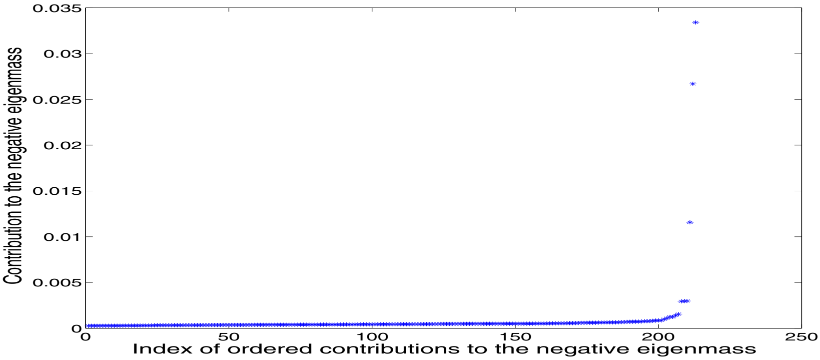


Fig. 4. Sorted each object's contribution to the negative eigenvalues at Protein

We plot the individual contribution to the negative eigenmass for the Protein dataset in Figure 4. This shows that the negative eigenvalues are caused by the non-Euclidean distances of just a few objects. The protein data is almost Euclidean with a very small negative eigenfraction value of 0.001. We have explored the effect of applying a leave one out nearest neighbor classifier to the dataset. When we edit out the effect of the outlier objects distances by adding a constant to the squared distances to the remaining objects, we obtain only a slightly smaller error rate of 0.47% compared to 1.9% for the original distances.

4 Conclusion

This paper discusses three possible sources of non-Euclidean behavior in dissimilarity data. We present three measures for analysing and determining the causes of negative eigenvalues in a non-Euclidean dissimilarity matrix. The three measures are based on distribution of the negative eigenvalues, and allow us to determine if the case is a) that data resides on a manifold, b)that the objects may be extended and c) that there is Gaussian noise.

Acknowledgments. Edwin R. Hancock is supported by the EU FET project SIMBAD and by a Royal Society Wolfson Research Merit Award.

References

1. Wilson, R., Hancock, E.: Spherical embedding and classification. In: SSPR (2010)
2. Duin, R.P.W., Pkekalska, E.: Non-Euclidean dissimilarities causes and informativeness. In: SSPR, pp. 324–333 (2010)
3. Pekalska, E., Harol, A., Duin, R., Spillmann, B., Bunke, H.: Non-Euclidean or Nonmetric Measures Can Be Informative. In: SSPR, pp. 871–880 (2006)

4. Lauba, J., Rothb, V., Buhmannb, J.M., Mllera, K.-R.: On the information and representation of non-Euclidean pairwise data. Pattern Recognition, 1815–1826 (2006)
5. Goldfarb, L.: A new approach to pattern recognition. Progress in Pattern Recognition, 241–402 (1985)
6. Sanfeliu, A., Fu, K.-S.: A Distance measure between attributed relational graphs for pattern recognition. IEEE Transactions on Systems, Man, and Cybernetics, 353–362 (1983)

Robust Model-Based Detection of Gable Roofs in Very-High-Resolution Aerial Images

Lykele Hazelhoff[1,2] and Peter H.N. de With[2]

[1] CycloMedia Technology B.V., The Netherlands
lhazelhoff@cyclomedia.com
[2] Eindhoven University of Technology, The Netherlands
p.h.n.de.with@tue.nl

Abstract. This paper describes an improved version of our system for robust detection of buildings with a gable roof in varying rural areas from very-high-resolution aerial images. The algorithm follows a custom-made design, extracting key features close to modeling, such as roof ridges and gutters, in order to allow a large freedom in roof appearances. It starts by detecting straight line-segments as roof-ridge hypotheses, and for each of them, the likely roof-gutter positions are estimated. Supervised classification is employed to select the optimal gutter pair and to reject unlikely detections. Afterwards, overlapping detections are merged. Experiments on a large dataset containing 220 images, covering different rural regions with significant variation in both building appearance and surroundings, show that the system is able to detect over 87% of the present buildings, thereby handling common distortions, such as occlusions by trees.

Keywords: Building detection, Object detection, Remote sensing.

1 Introduction

Very-high-resolution aerial images are captured from The Netherlands at a yearly basis, providing a recent overview of the country infrastructure. Updating of civil community databases based on aerial images is time consuming when performed manually, leading to a demand for automated interpretation of these images. This is of particular interest in rural areas, since these cover a widespread area together with a low population density, increasing the cost per citizen. Since buildings are dominant features in those images, accurate extraction of their locations is important. However, accurate and large-scale detection of buildings in aerial images is a complicated problem, since buildings vary considerably in appearance, may feature complex compositions and occlusions by trees occur frequently. Besides this, large variations also exist in the source data due to varying capturing conditions, illumination circumstances and sensor differences. This causes large variations in both visual appearance and statistical properties. Therefore, development of a building-detection solution that is able to handle these issues is a rather complicated problem.

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 598–605, 2011.

In literature, localization of buildings in remote sensing data is researched for decades, based on many different types of source data, including satellite and aerial images, lidar data, etc. As the information content of these sources varies from $2D$ grayscale images to $3D$ structural information, the proposals vary from building localization to automatic $3D$ city reconstruction. However, our country-covering datasets only contain color aerial images, disabling the use of any $3D$ information. Relevant proposals in literature for detection of buildings in *single* color images are e.g. based on image segmentation [1]. There, the input image is segmented using a range of parameters, resulting in multi-layer segmented images. This is analyzed with a tree structure by means of rooftop constraints. Potential buildings are evaluated relying on shadow information and a fixed height. Other strategies include searching of closed loops [2]. In this work, after mean-shift segmentation, edge pixels that form closed loops are converted into polygons. These polygons are used to deduce the building shape. Nosrati *et al.* [3] apply dynamic programming to line intersections for searching of closed loops. Both rely on the visibility of the outer edges, which possibly disables detection of buildings in low-contrast situations. A combination of multiple information sources is followed by Jin *et al.* [4], where three different detectors, focusing at structural, contextual and spectral information, are applied for localization of buildings in high-resolution satellite images. They report a significant increase in detection performance by combining these detectors. Benedek *et al.* [5] apply a probabilistic approach for building extraction, where building footprints are represented by combinations of rectangular segments. The optimal building configuration is retrieved based on a global optimization process.

Although many proposals report accurate results, the test data is often limited in both numbers and in-set variety. In contrast, we focus on robustness to the above-described variations, and we have described a novel, specific algorithm for localization of buildings with a gable roof in rural areas [6]. Whereas we aim at applying the algorithm at a large-scale database, we have designed the algorithm for robustness against both variations in building appearance, variations in the source data and commonly occurring distortions, such as overhanging trees. These constraints have guided us to design a more generic system, aiming at an high overall score, instead of a very high score in a specific situation, which is a different approach compared to many proposals in literature. In this paper, we describe an improved version of the original algorithm, which still operates under the above-mentioned conditions. Next to this, we will also present new and more accurate results obtained with this system, where we used an extended version of our test set, now containing more images and captured in three different years.

2 Algorithm Description

2.1 Preprocessing

Figure 1 portrays the schematic overview of our algorithm, based on [6].

Prior to image analysis, the image is segmented using a region-segmentation procedure similar to [7]. Vegetated and large, uniform farmland areas are then

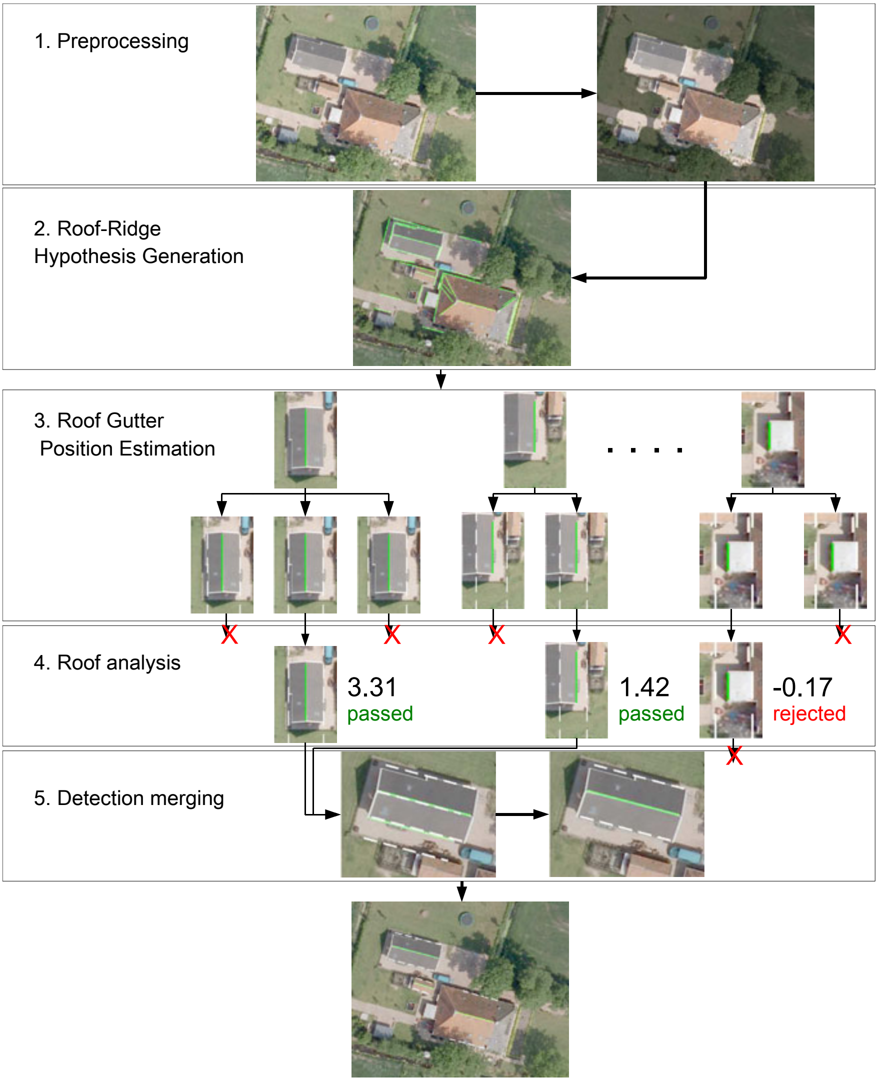


Fig. 1. Schematic overview of our building detection system

discarded using vegetation detection and a minimum size constraint of 500 m^2. Pixels (represented by RGB values) are marked as vegetation when they satisfy the following empirically determined rule:

$$G \geq \min\left(1.175 \cdot B\ ,\ 0.975 \cdot R\right). \tag{1}$$

2.2 Roof-Ridge Hypothesis Generation

Within the accepted regions of interest, hypotheses of occurring roof ridges are generated by detection of straight line segments. For this, a Canny edge detector is applied on a single-channel version O of the input RGB image, where the employed color transformation is chosen such that clear transitions are expected along roof ridges. Since roofs are usually either red or gray, and often contain a shadowed roof side, the following transformation is selected:

$$O = 0.5 \cdot (R + B)\,. \tag{2}$$

Connected edge pixels forming straight lines are extracted based on a technique described in [8], resulting in a set of line pieces. To enable by-passing of line-interrupting objects, like chimneys, individual line pieces having a position and orientation such that they jointly form a straight line are combined when they are located near the same region segment.

2.3 Roof Gutter Position Estimation

For all hypotheses of occurring roof ridges, the set of likely roof-gutter positions is estimated for both sides of the ridge. This results in a set of hypotheses of the roof configuration, given by each combination of gutter positions at both ridge sides, as displayed in Box 3 Fig. 1.

To identify the likely gutter positions, two rectangular regions located parallel to the ridge are analyzed, where the region size is inferred from training data. For ease of description, we assume that the ridge is oriented vertically. Within the regions, vertical Sobel filters are applied to both the red and blue color channels, since roofs are usually red or gray. Per color channel and for each column, the resulting edge energy is sorted by magnitude, and divided into 4 groups, each containing the sum of the represented 25% of the pixels. The resulting signal matrix **ES** containing 8 features per column i is represented with a Gaussian model. This model is trained from the first 10 columns, i.e. the columns closest to the roof ridge. The distance towards the center of the ellipsoid is calculated, which is proportional to:

$$D\left(\mathbf{ES}\left(i\right)\right) \sim e^{-\frac{1}{2}\left([\mathbf{ES}(i)-\boldsymbol{\mu}_{ES}]^T \mathbf{C}_{ES}^{-1}[\mathbf{ES}(i)-\boldsymbol{\mu}_{ES}]\right)}, \tag{3}$$

where $\boldsymbol{\mu}_{ES}$ and $\mathbf{C}_{ES}$ denote the mean and covariance matrix extracted from the 10 training samples. Column i with $D\left(\mathbf{ES}\left(i\right)\right)$ satifying (1) local maximality, (2) higher than twice the running average, and (3) larger than 66% of the running maximum, is identified as roof-gutter position.

2.4 Roof Analysis

Each roof-configuration hypothesis is validated, where at first infeasible candidates are rejected, and second, machine learning is employed to classify each remaining configuration between roof half and non roof-half. This results in a likelihood value for each configuration, used to select the optimal roof configuration for each ridge hypothesis and to discard low-valued detections.

The first step aims at rejecting very unlikely configurations by checking a number of loose constraints on size, aspect ratio, shadow profile and vegetation coverage. This step exploits physical properties of gable roofs, and is described in detail [6].

All remaining configurations are subject to classification based on supervised classification. For each of both roof halves, the following features are extracted.

1. *Roof-Gutter Edge Orientation Histogram*: a gradient orientation histogram is extracted along the gutters, indicating the deviation in orientation w.r.t. the ridge orientation. This histogram contains 6 bins representing 15° each.
2. *Segmentation Features*: The five largest clusters covering each roof half are expressed as a percentage of the total amount of roof-half pixels.
3. *Color Similarity Histogram*: patch-based similarity analysis is employed to investigate roof similarities. For each patch, the most similar patch is searched (with a minimum distance constraint), and after sorting, the values for which 20%, 40%, etc. of the pixels are lower, are extracted.

Next to these features, also the ridge length, aspect ratio and ratio of roof-half widths are extracted. The total feature vector consists of 51 features, containing the 3 global features, 2×16 features for each roof half, appended by their sum. Note that during the training phase, we have also included the samples with reversed roof halves. Classification is performed using a Support Vector Machine (SVM) with radial basis function kernel, which outputs the distance towards the decision bound, as a kind of likelihood information. For each ridge hypothesis, the configuration with the highest value is selected as optimal configuration, while configurations with a low SVM output value are discarded.

2.5 Detection Merging

Each physical roof may contain several detections, where at most one detection corresponds to the actual roof ridge, as shown in Box 5 in Fig. 1. Therefore, overlapping detections are fused using supervised classification, where we have found that multiple classifiers rejecting detections in specific situations outperform a single, large classifier. As a result, we have installed multiple linear SVM classifiers, each of them set up to analyze a specific situation. These situations include differences in red profile, compliance with the expected shadow profile, detections with complete overlap, etc. For each situation, we extract relevant features, and append this by features representing the differences in likelihood, length and angle between the ridges. The continuous output of each classifier is subject to a threshold such that at most 0.5% of the positive training samples are rejected. This threshold setting gives a trade-off between detections and false alarms.

Note that each classifier fuses a small fraction of the overlapping detections. In some cases, the optimal fusion process does not satisfy our generic principles, so that not all overlapping detections are fused. Even then, the amount of detections is reduced significantly.

3 Experiments and Results

3.1 Dataset Description and Test Setup

The algorithm is tested on 220 ortho-rectified aerial images, captured in 2008, 2009 and 2010 during spring and fall time. All images are normalized such that each pixel corresponds to $0.1\,m$ on the ground plane. The set contains $1,523$ buildings, including houses, garages, barns and small sheds, which are all marked manually for evaluation purposes. The set is randomly divided in two equal parts, denoted as $SetA$ and $SetB$, where for each test, the other set is used for training.

The performance of our system is analyzed in two ways. First, we have evaluated the detection accuracy of the system, i.e. how accurate our algorithm indicates the presence of a gable roof. Second, we have analyzed the accuracy of the estimated building dimensions, i.e. how well the algorithm indicates the building size. These two aspects are separated since we concentrate on detection of buildings, where we consider the size estimate as a byproduct.

The detection performance is assessed using the following metrics:

- Recall: $R = \frac{TP}{TP+FN} \cdot 100\%$,
- Precision: $P = \frac{TP}{TP+FP} \cdot 100\%$.

We count a detection as a *True Positive* (detected building) when it overlaps the roof ridge for at least 51%, all other detections are treated as *False Positives* (falsely detected buildings). Since we aim at localizing the somewhat larger buildings in rural areas, we have neglected buildings smaller than 5×3 meters based on a minimum roof-ridge length and building width. These minimum sizes are chosen such that typical side-buildings, like e.g. garages, are still detected.

3.2 Results

The recall-precision curves for both sets are shown in Fig. 2. As follows from this graph, our system is capable of detecting over 87% of the target buildings, where for a detection rate of 80%, around 77% of the detections are correct. For this specific working point, we have analyzed the performance in more detail. Buildings are missed due to large occlusions by trees, or shadows from trees on the roof. We have already reported in [6] that about 15% of the buildings are covered by trees to a certain extent, where we add that about 2% of the buildings have at least one roof half covered completely by trees (as in e.g. Fig. 3(h)). In numerous cases, we still find a detection on the outer building border, but these are counted as false detections. Other causes of misdetections are e.g. low-contrast roof ridges, which especially occur at completely black roofs. Such special cases may be handled by additional detectors, as e.g. applied in [4]. False detections are mainly located at the outer building borders of missed buildings and on gable-roof-like objects, including road segments. For both categories, the SVM output is relative low, but not smaller than some detected buildings. Fig. 3 displays some examples of correct detections, false detections and missed buildings; more examples can be found on our website[1].

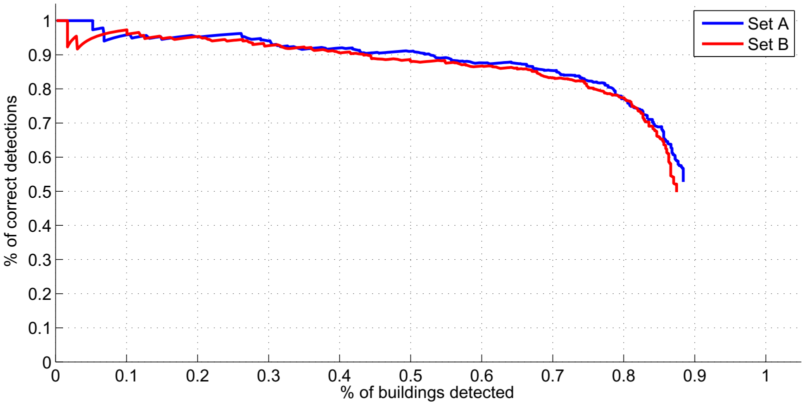


Fig. 2. Recall-precision curve for our system, shown for both *setA* and *setB*

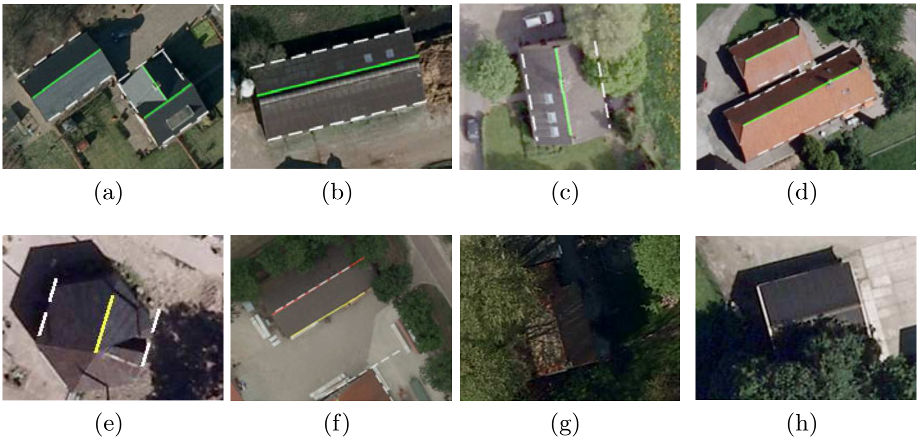


Fig. 3. Examples of correct detections (a)-(d), false detection (e)-(f) and missed building (f)-(h). The estimated ridge positions are drawn green (TP), yellow (FP) or red (FN); corresponding gutter positions are drawn dotted white .

We have also assessed the modeling capabilities of our algorithm by comparing the estimated length and width against the ground-truth dimensions. The found roof ridge is often smaller than the real roof length, due to chimneys and roofs where the ridge ends in a triangular shape, as in e.g. Fig. 3(d). The main cause for inaccurately found building widths are low-contrast gutters. Quantitatively, 80.2% of the detected ridges deviate less than 10% from the ground-truth length, 83.7% of the estimated widths deviate less than 10% and 61.8% of the detections have an area deviation smaller than 10%. We consider this as a reason-

[1] See http://vca.ele.tue.nl/demos/buildingdetection/index.html.

ably accurate result as our system is primarily targeting at localizing buildings with a diverse test set, where numerous buildings are overlapped by trees.

4 Conclusions and Future Work

We have presented an improved version of our algorithm for detection of gable roofs in very-high-resolution aerial images, focusing on rural areas. Our system aims at detecting buildings with a high robustness in a generic way, and is therefore designed with special attention to the large existing variations in both building appearance and source data. This deviates from the common practice to pursue high scores in a specific number of cases. A specific novel aspect of our algorithm is the example-based detection-merging step, based on multiple linear SVM classifiers. Tests on a diverse dataset, containing numerous geographical locations, have shown that our system is able to localize over 87% of the buildings larger than $5 \times 3\,m$, where for a recall of 80%, a precision of around 77% is obtained. The reported system is able to detect around 8% more buildings than our previously described system. Considering that our test set contains large variations in both source data and target objects, and no height information is incorporated in the detection process, we consider this as an accurate result. Improvement of the modeling capabilities is part of our future work.

References

1. Izadi, M., Saeed, P.: Automatic building detection in aerial images using a hierarchical feature based image segmentation. In: 20th International Conference on Pattern Recognition, pp. 472–475 (2010)
2. Ok, A.O.: Automated description of 2-d building boundaries from a single color aerial ortho-image. In: Proceedings of ISPRS, High Res. Earth Imag. for Geospat. Inf., pp. 1417–1420 (2009)
3. Nosrati, M.S., Saeedi, P.: A novel approach for polygonal rooftop detection in satellite/aerial imageries. In: Int. Conf. on Image Processing (ICIP), pp. 1709–1712 (2009)
4. Jin, X., Davis, C.H.: Automated building extraction from high-resolution satellite imagery in urban areas using structural, contextual, and spectral information. EURASIP Journal on Applied Signal Processing, 2196–2206 (2005)
5. Benedek, C., Descombes, X., Zerubia, J.: Building detection in a single remotely sensed image with a point process of rectangles. In: 20th International Conference on Pattern Recognition (2010)
6. Hazelhoff, L., De With, P.: Localizations of buildings with a gable roof in very-high-resolution aerial images. In: Proceedings of IS&T SIE Electronic Imaging, Visual Information Processing and Communication II (2011)
7. Comanicu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 603–619 (2002)
8. Guru, D.S., Shekar, B.H., Nagabhushan, P.: A simple and robust line detection algorithm based on small eigenvalue analysis. Pattern Recognition Letters 25, 1–13 (2004)

Author Index