Jorge S. Marques
Nicolás Pérez de la Blanca
Pedro Pina (Eds.)

LNCS 3523

Pattern Recognition and Image Analysis

Second Iberian Conference, IbPRIA 2005
Estoril, Portugal, June 2005
Proceedings, Part II

2 Part II

Springer

Jorge S. Marques Nicolás Pérez de la Blanca
Pedro Pina (Eds.)

Pattern Recognition and Image Analysis

Second Iberian Conference, IbPRIA 2005
Estoril, Portugal, June 7-9, 2005
Proceedings, Part II

Springer

Volume Editors

Jorge S. Marques
Instituto Superior Técnico, ISR
Torre Norte, Av. Rovisco Pais, 1049-001, Lisboa, Portugal
E-mail: jsm@isr.ist.utl.pt

Nicolás Pérez de la Blanca
Universidad de Granada, ETSI Informática
Departamento de Ciencias de la Computacíon e Inteligencia Artificial
Periodista Daniel Saucedo Aranda s/n, 18071 Granada, Spain
E-mail: nicolas@ugr.es

Pedro Pina
Instituto Superior Técnico, CVRM
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
E-mail: ppina@alfa.ist.utl.pt

Library of Congress Control Number: 2005926832

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

ISSN 0302-9743
ISBN-10 3-540-26154-0 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-26154-4 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Typesetting: Camera-ready by author, data conversion by Olgun Computergrafik
Printed on acid-free paper SPIN: 11492542 06/3142 5 4 3 2 1 0

Preface

IbPRIA 2005 (Iberian Conference on Pattern Recognition and Image Analysis) was the second of a series of conferences jointly organized every two years by the Portuguese and Spanish Associations for Pattern Recognition (APRP, AERFAI), with the support of the International Association for Pattern Recognition (IAPR).

This year, IbPRIA was hosted by the Institute for Systems and Robotics and the Geo-systems Center of the Instituto Superior Técnico and it was held in Estoril, Portugal. It provided the opportunity to bring together researchers from all over the world to discuss some of the most recent advances in pattern recognition and all areas of video, image and signal processing.

There was a very positive response to the Call for Papers for IbPRIA 2005. We received 292 full papers from 38 countries and 170 were accepted for presentation at the conference. The high quality of the scientific program of IbPRIA 2005 was due first to the authors who submitted excellent contributions and second to the dedicated collaboration of the international Program Committee and the other researchers who reviewed the papers. Each paper was reviewed by two reviewers, in a blind process. We would like to thank all the authors for submitting their contributions and for sharing their research activities. We are particularly indebted to the Program Committee members and to all the reviewers for their precious evaluations, which permitted us to set up this publication.

We were also very pleased to benefit from the participation of the invited speakers Prof. David Lowe, University of British Columbia (Canada), Prof. Wiro Niessen, University of Utrecht (The Netherlands) and Prof. Isidore Rigoutsos, IBM Watson Research Center (USA). We would like to express our sincere gratitude to these world-renowned experts.

We would like to thank Prof. João Sanches and Prof. João Paulo Costeira of the Organizing Committee, in particular for the management of the Web page and the submission system software.

Finally, we were very pleased to welcome all the participants who attended IbPRIA 2005. We are looking forward to meeting you at the next edition of IbPRIA, in Spain in 2007.

Estoril, June 2005

Jorge S. Marques
Nicolás Pérez de la Blanca
Pedro Pina

Conference Chairs

Jorge S. Marques	Instituto Superior Técnico
Nicolás Pérez de la Blanca	University of Granada
Pedro Pina	Instituto Superior Técnico

Organizing Committee

João M. Sanches	Instituto Superior Técnico
João Paulo Costeira	Instituto Superior Técnico

Invited Speakers

David Lowe	University of British Columbia, Canada
Wiro Niessen	University of Utrecht, The Netherlands
Isidore Rigoutsos	IBM Watson Research Center, USA

Supported by

Fundação Oriente, Lisbon
Fundação para a Ciência e Tecnologia
HP Portugal
Institute for Systems and Robotics, Lisbon
International Association for Pattern Recognition

Program Committee

Jake Aggarwal	University of Texas, USA
Hélder Araújo	University of Coimbra, Portugal
José Benedi	Polytechnic University of Valencia, Spain
Isabelle Bloch	ENST, France
Hervé Bourlard	EPFL, Switzerland
Patrick Bouthemy	IRISA, France
Horst Bunke	University of Bern, Switzerland
Aurélio Campilho	University of Porto, Portugal
Gilles Celeux	Université Paris-Sud, France
Luigi Cordella	University of Naples, Italy
Alberto Del Bimbo	University of Florence, Italy
Hervé Delinguette	INRIA, France
Rachid Deriche	INRIA, France
José Dias	Instituto Superior Técnico, Portugal
Robert Duin	University of Delft, The Netherlands
Mário Figueiredo	Instituto Superior Técnico, Portugal
Ana Fred	Instituto Superior Técnico, Portugal
Andrew Gee	University of Cambridge, UK
Mohamed Kamel	University of Waterloo, Canada
Aggelos Katsaggelos	Northwestern University, USA
Joseph Kittler	University of Surrey, UK
Seong-Whan Lee	University of Korea, Korea
Ana Mendonça	University of Porto, Portugal
Hermann Ney	University of Aachen, Germany
Wiro Niessen	University of Utrecht, The Netherlands
Francisco Perales	Universitat de les Illes Balears, Spain
Maria Petrou	University of Surrey, UK
Armando Pinho	University of Aveiro, Portugal
Ioannis Pitas	University of Thessaloniki, Greece
Filiberto Pla	University Jaume I, Spain
Richard Prager	University of Cambridge, UK
José Principe	University of Florida, USA
Ian Reid	University of Oxford, UK
Gabriella Sanniti di Baja	Istituto di Cibernética, Italy
Beatriz Santos	University of Aveiro, Portugal
José Santos-Victor	Instituto Superior Técnico, Portugal
Joan Serrat	Universitat Autònoma de Barcelona, Spain
Yoshiaki Shirai	Osaka University, Japan
Pierre Soille	Joint Research Centre, Italy
Karl Tombre	LORIA, France
M. Ines Torres	University of the Basque Country, Spain
Emanuele Trucco	Heriot-Watt University, UK
Alessandro Verri	University of Genoa, Italy
Max Viergever	University of Utrecht, The Netherlands
Joachim Weickert	Saarland University, Germany

Reviewers

Arnaldo Abrantes
Luís Alexandre
René Alquézar
Juan Carlos Amengual
Teresa Barata
Jorge Barbosa
Jorge Batista
Luis Baumela
Alexandre Bernardino
Javier Binefa
Hans Du Buf
Francisco Casacuberta
Miguel Velhote Correia
Paulo Correia
João P. Costeira
Jose Manuel Fuertes
José Gaspar
Edwin Hancock
Francisco Mario Hernández
Arturo De La Escalera Hueso
Jose Manuel Iñesta
Alfons Juan
João Miranda Lemos
Manuel Lucena Lopez
Javier Lorenzo
Maria Angeles Lopez Malo
Elisa Martínez Marroquín
Jesus Chamorro Martinez
Eduard Montseny Masip
Nicolás Guil Mata
Luisa Micó
Rafael Molina
Ramón A. Mollineda
Jacinto Nascimento
Jesus Ariel Carrasco Ochoa
Paulo Oliveira
António Guedes Oliveira
Arlindo Oliveira
Antonio Adan Oliver
José Oncina
Roberto Paredes
Antonio Miguel Peinado
Fernando Pereira
André Puga
Francesc Josep Ferri Rabasa
Juan Mendez Rodriguez
Antoni Grau Saldes
João M. Sanches
José Salvador Sánchez
Modesto Castrillon Santana
José Ruiz Shulcloper
Jorge Alves Silva
Margarida Silveira
António Jorge Sousa
João M. Sousa
João Tavares
António J.S. Teixeira
Ana Maria Tomé
Jose Ramon Fernandez Vidal
Enrique Vidal
Juan Jose Villanueva
Jordi Vitrià

Table of Contents, Part II

I Statistical Pattern Recognition

II Syntactical Pattern Recognition

III Image Analysis

IV Document Analysis

V Bioinformatics

VI Medical Imaging

VII Biometrics

VIII Speech Recognition

IX Natural Language Analysis

X Applications

Table of Contents, Part I

I Computer Vision

II Shape and Matching

III Image and Video Processing

IV Image and Video Coding

V Face Recognition

VI Human Activity Analysis

VII Surveillance

VIII Robotics

IX Hardware Architectures

Part I

Statistical Pattern Recognition

Testing Some Improvements of the Fukunaga and Narendra's Fast Nearest Neighbour Search Algorithm in a Spelling Task

Eva Gómez-Ballester, Luisa Micó, and Jose Oncina*

Dept. Lenguajes y Sistemas Informáticos,
Universidad de Alicante, E-03071 Alicante, Spain
{eva,mico,oncina}@dlsi.ua.es

Abstract. Nearest neighbour search is one of the most simple and used technique in Pattern Recognition.

One of the most known fast nearest neighbour algorithms was proposed by Fukunaga and Narendra. The algorithm builds a tree in preprocess time that is traversed on search time using some elimination rules to avoid its full exploration.

This paper tests two new types of improvements in a real data environment, a spelling task. The first improvement is a new (and faster to build) type of tree, and the second is the introduction of two new elimination rules.

Both techniques, even taken independently, reduce significantly both: the number of distance computations and the search time expended to find the nearest neighbour.

1 Introduction

The Nearest Neighbour Search method consists on finding the nearest point of a set to a given test point using a distance function [3].

To avoid the exhaustive search many effective algorithms have been developed [1]. Although some of such algorithms as K-dtrees, R-trees, etc. depend on the way the points are represented (vectors usually), in this paper we are going to focus on algorithms that does not make any assumption on the way the points are represented making them suitable to work in any metric space.

The most popular and refereed algorithm of such type was proposed by Fukunaga and Narendra (FNA) [4]. Although some recently proposed algorithms are more efficient, the FNA is a basic reference in the literature and in the development of new rules to improve the main steps of the algorithm that can be easily extended to other tree based algorithms [2, 6, 9].

* The authors thank the Spanish CICyT for partial support of this work through project TIC2003–08496-CO4, GV04B-541, GV04B-631, and the IST Programme of the Europen Community, under the PASCAL Network of Excelence, IST–2002-506778. This publication only reflects the authors' view.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 3–10, 2005.

Recently we proposed some improvements in this algorithm [11] that reduce significantly the number of distance computations. However, those improvements were tested only with data represented in a vector space. In this work the algorithm is checked in a spelling task where the points are represented by strings and the distance function is the edit distance. Also we compare our proposal with Kalantari and McDonalds method as well as with FNA.

2 The Fukunaga and Narendra Algorithm

The FNA is a fast search method that uses a tree structure. Each node p of the tree represents a group of points S_p, and is characterised by a point $M_p \in S_p$, (the representative of the group S_p), and its distance R_p of the farthest point in the set (the radius of the node). The tree is built using recursively the c-means clustering algorithm.

When a new test point x is given, its nearest neighbour n is found in the tree using a first-depth strategy. Among the nodes at the same level, the node with a smaller distance $d(x, M_p)$ is searched earlier. In order to avoid the exploration of some branches of the tree, the FNA uses a prune rule.

Rule: if n is the nearest neighbour to x up to the moment, no $y \in S_p$ can be the nearest neighbour to x if

$$d(x, n) + R_p < d(x, M_p)$$

This rule will be referenced as the Fukunaga and Narendra's Rule (FNR) (see fig. 1 for a graphical interpretation).

The FNA defines another rule in order to avoid some distance computations in the leaves of the tree. In this work only binary trees with one point on the leaves are considered. On such case the rule related to leaf nodes becomes a special case of the FNR and will not be considered on the following.

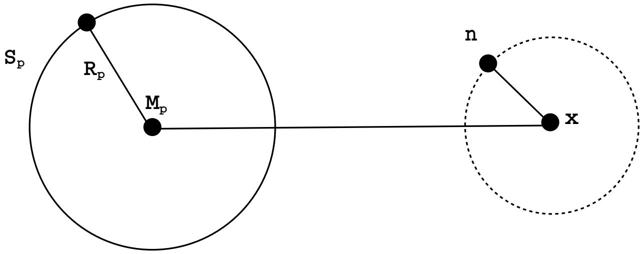


Fig. 1. Original elimination rule used in the algorithm of Fukunaga and Narendra (FNR).

3 The Search Tree

In previous works [11] some approximations were developed as an alternative to the use of the c-means algorithm on the construction of the tree. The best behaviour was obtained by the method called *Most Distant from the Father tree* (MDF). In this work this strategy is compared with c-means strategy[1] and with the incremental strategy to build the tree proposed by Kalantari and McDonalds [5], since this last strategy builds a binary tree similar to ours. Given a set of points, the MDF strategy consists on

- randomly select a point as the representative of the root node;
- in the following level, use as representative of the left node the representative of the father node. The representative of the right node is the farthest point among all the points belonging to the father node;
- classify the rest of the prototypes in the node of their nearest representative;
- recursively repeat the process until each leaf node has only one point, the representative.

This strategy reduces the computation of some distances in the search procedure as the representative of the left node is the same than the representative of its father. Each time a expansion of the node is necessary, only one new distance should be computed. Note that the construction of this tree is much faster than the construction of the FN tree where the c-means algorithm is used recursively.

While in the MDF method the average time complexity is $O(n \log(n))$, in the case that c-means algorithm is used, the average complexity is $O(n^2 \log(n))$.

4 The New Elimination Rules

In the proposed rules, to eliminate a node ℓ, also information related with the sibling node r is used.

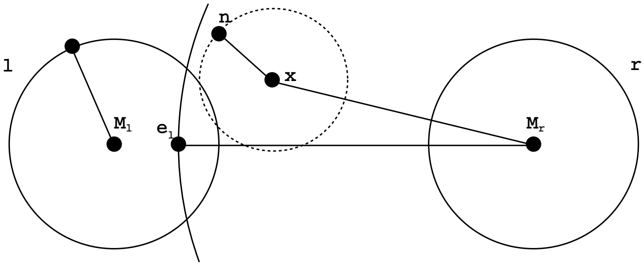


Fig. 2. Sibling based rule (SBR).

[1] As data are strings, the mean of a set of points can't be obtained. In this case the median of the set is used (c-medians).

4.1 The Sibling Based Rule (SBR)

The main idea of this rule is to use the distance from the representative to the nearest prototype of the sibling node. If this distance is too big, the sibling node can be safely ignored. Kamgar-Parsi and Kanal [10] proposed, for the FN algorithm, a similar rule (KKR), but the distance from the mean to the nearest prototype in the node was used. Note that in our case the representative is always a prototype in the node, then this distance is always zero. Moreover, in our case the rule allows the pruning of the sibling node, in the KKR case is the own node that can be pruned.
A first proposal requires that each node r stores the distance between the representative of the node, M_r, and the nearest point, e_ℓ, in the sibling node ℓ.

Definition 1. ***Definition of SBR:*** *given a node r, a test sample x, an actual nearest neighbour n, and the nearest point to the representative of the sibling node ℓ, e_ℓ, the node ℓ can be pruned if the following condition is fulfil (fig. 2):*

$$d(M_r, e_\ell) > d(M_r, x) + d(x, n)$$

Unlike the FNR, SBR can be applied to eliminate node ℓ without compute $d(M_\ell, x)$. That permits to avoid some distance computations in the search procedure.

4.2 Generalised Rule (GR)

This rule is an iterated combination of the FNR and the SBR. Given a node ℓ, a set of points $\{t_i\}$ is defined in the following way:

$$\begin{aligned} G_1 &= S_\ell \\ t_i &= \mathrm{argmax}_{p \in G_i} d(p, M_\ell) \\ G_{i+1} &= \{p \in G_i : d(p, M_r) < d(t_i, M_r)\} \end{aligned}$$

where M_r is the representative of the sibling node r, and G_i are auxiliary sets of points needed in the definition (fig. 3). In preprocessing time, the distances $d(M_r, t_i)$ are stored in each node ℓ. In the same way, this process is repeated for the sibling node.

Definition 2. **Definition of GR:** *given two sibling nodes ℓ and r, a test sample x, an actual nearest neighbour n, and the list of point $t_1, t_2, \ldots, t_s$, the node ℓ can be pruned if there is an integer i such that:*

$$d(M_r, t_i) \geq d(M_r, x) + d(x, n) \tag{1}$$

$$d(M_\ell, t_{i+1}) \leq d(M_\ell, x) - d(x, n) \tag{2}$$

Cases $i = 0$ and $i = s$ are also included not considering equations (1) or (2) respectively. Note that condition (1) is equivalent to SBR rule when $i = s$ and condition (2) is equivalent to FNR rule when $i = 0$.

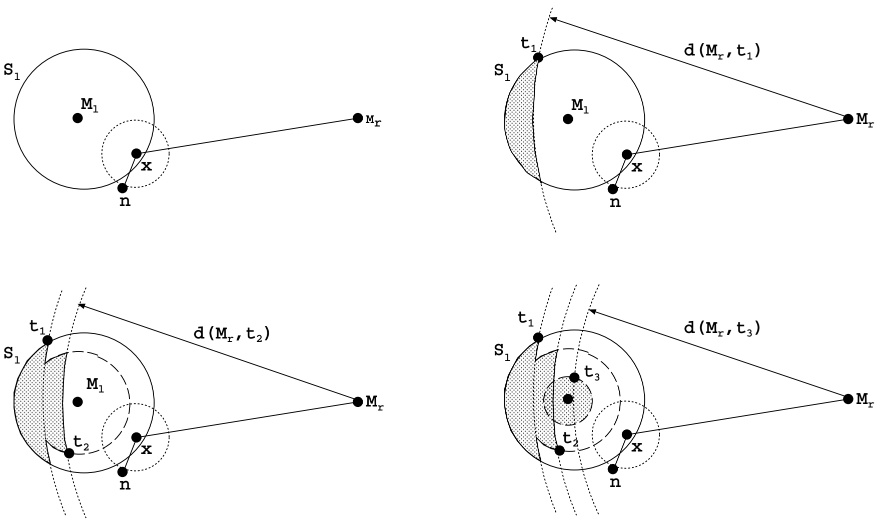


Fig. 3. Generalised rule (GR).

5 Experiments

To show the performance of the algorithm some tests were carried out on a spelling task. A database of 38000 words of a Spanish dictionary was used. The input test of the speller was simulated distorting the words by means of random insertion, deletion and substitution operations over the words in the original dictionary. The edit distance was used to compare the words.

Increasing size of dictionaries (from 1000 to 38000, 9 different sizes) was obtained extracting randomly words of the whole dictionary. The test points were 1000 distorted words obtained from randomly selected dictionary words. To obtain reliable results the experiments were repeated 10 times. The averages and the standard deviations are showed on the plots. The distance computation is referenced per test point, and the search time per test set.

In order to study the contribution of the elimination rules FNR, FNR+SBR and GR a first set of experiments were carried out using the original *c*-means tree construction of the FN algorithm (fig. 4).

As was expected, the addition of the SBR reduces slightly the number of distance computations and the search time, but GR reduces them drastically (to less than one half).

A second set of experiments were carried out in order to compare the MDF method of tree construction to the original *c*-means method (using the FNR) and to the incremental strategy proposed by Kalantari and Mc-Donalds (KM).

Figure 5 illustrates the average number of distance computations and the search time using the *c*-means, KM and MDF tree construction methods. It can be observed that the MDF reduces to less than one half the number of distance computation and the search time.

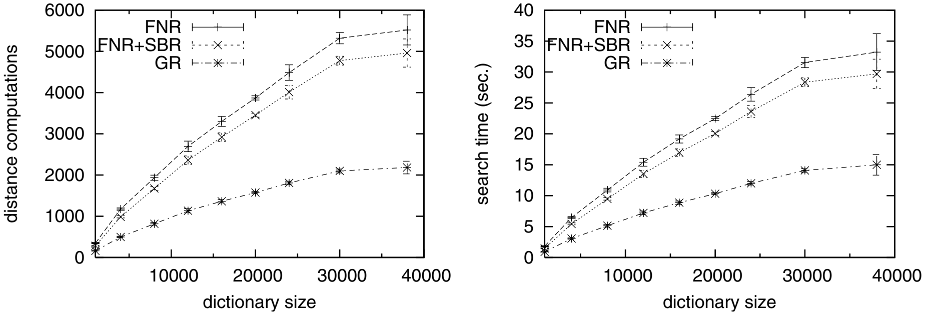


Fig. 4. Comparison of FNR, FNR+SBR and GR elimination rules using the c-medians tree construction.

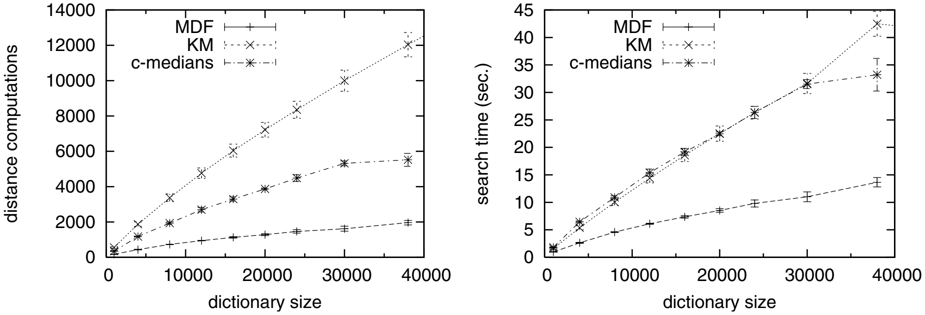


Fig. 5. Comparison of *c*-medians, KM and MDF methods to build the tree using the FNR elimination rule.

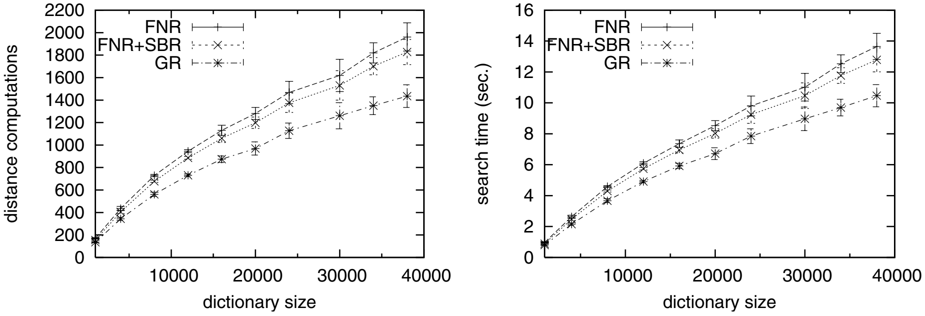


Fig. 6. Comparison of FNR, FNR+SBR and GR using a tree constructed with MDF.

Once stated that the MDF method is superior that the *c*-means method, a third set of experiments were carried out in order to study the contribution of FNR, FNR+SGR and GR with a tree constructed following MDF method.

As figure 6 shows, the number of distance computations and the search time decrease using the new rules although now the reductions are not so impressive that in the previous cases.

Nevertheless, comparing the distance computations and the search time of the original algorithm (fig. 4, FNR) to the algorithm using GR and MDF (fig. 6,GR), it can be observed that applying both techniques at the same time the distance computations and the search time can be reduced one third.

6 Conclusions

In this work two improvements of the Fukunaga and Narendra fast nearest neighbour algorithm was tested in a spelling correction task.

The first improvement is a new method to build the decision tree used in the FN algorithm. On one hand, to build the tree with this method it is much faster than with the original one and, on the other hand, the use of this method reduces the number of distance computations and the search time to one half in our experiments. The use of the KM way to build the tree increases the number of distance computations even more than with the original method. The second modification is the introduction of two new elimination rules. The use of the GR rule reduces to one half the number of distance computations and the search time. Both improvements can be applied together reaching reductions to one third in the distance computations and the search time.

On this work the generalised elimination rule was implemented in a quite naive way by means of an iterative procedure. Now we are interested in implementing this rule using a more adequate algorithm to obtain further search time reductions.

We believe that these approximations can be extended to other nearest neighbour search algorithms based on a tree structure.

References

1. Belur V. Dasarathy: Nearest Neighbour(NN) Norms: NN Pattern Classification Techniques. IEEE Computer Society Press (1991).
2. Chen, Y.S., Hung, Y.P., Fuh, C.S.: Fast Algorithm for Nearest Neighbour Search Based on a Lower Bound Tree. Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada, (2001), **1** 446–453
3. Duda, R., Hart, P.: Pattern Classification and Scene Analysis. Wiley (1973)
4. Fukunaga, K., Narendra, M.: A branch and bound algorithm for computing k–nearest neighbours. IEEE Trans. Computing (1975) **24** 750–753
5. Kalantari, I., McDonald, G.: A data structure and an algorithm for the nearest point problem. IEEE Trans. Software Engineering (1983) **9** 631–634
6. Micó, L., Oncina, J., Carrasco, R.C.: A fast branch and bound nearest neighbour classifier in metric spaces. Pattern Recognition Letters (1996) **17** 731–739
7. Alinat, P.: Periodic progress report 4, ROARS project ESPRIT II - Number 5516. Thomson Technical Report TS ASM 93/S/EGS/NC/079 (1993)
8. S. Omachi, H. Aso: A fast algorithm for a k-NN classifier based on branch and bound method and computational quantity estimation. Systems and Computers in Japan, vol. 31, no 6, pp. 1-9 (2000).
9. Geofrey I. Webb: OPUS: An Efficient Admissible Algorithm for Unordered Search. Journal of Artificial Intelligence Research 3 (1995) 431-465.

10. Kamgar-Parsi, B., Kanal, L.: An improved branch and bound algorithm for computing k-nearest neighbors. Pattern Recognition Letters (1985) **3** 7–12
11. E. Gómez-Ballester, L. Micó, J. Oncina: Some improvements in tree based nearest neighbour algorithms. Progress in Pattern Recognition, Speech and Image Analysis. Proceedings of the 8th Iberoamerican Congress on Pattern Recognition. Havana, Cuba. Springer Verlag, Lecture notes in Artificial Intelligence, pp. 456–46 (2003)

Solving Particularization with Supervised Clustering Competition Scheme

Oriol Pujol and Petia Radeva

Computer Vision Center, Campus UAB, 08193 (Bellaterra) Barcelona, Spain
oriol@cvc.uab.es

Abstract. The process of mixing labelled and unlabelled data is being recently studied in semi-supervision techniques. However, this is not the only scenario in which mixture of labelled and unlabelled data can be done. In this paper we propose a new problem we have called particularization and a way to solve it. We also propose a new technique for mixing labelled and unlabelled data. This technique relies in the combination of supervised and unsupervised processes competing for the classification of each data point. Encouraging results on improving the classification outcome are obtained on MNIST database.

1 Introduction

The process of mixing labelled with unlabelled data to achieve classification improvement is being recently addressed by the community in the form of semi-supervision processes. This is a general denomination for a recently novel problem of pattern recognition based on the improvement of classification performance in the presence of very few labelled data. This line of work become very active since several authors point out the beneficial effects that unlabelled data can have [4–7].

However this is not the only scenario in which we can apply this mixture. Consider the following examples: Imagine a problem of handwritten character recognition, in which we know that all the characters in the document are written by the same author; consider an architect drawing a technical plane with its own symbols; a medical application in which we know that the data we need to classify proceeds from a single patient; the problem of face recognition, with a significative data set representing what a face is, and a huge set of data representing what a face is not. In all these problems there is a common fact, the knowledge of the fact that our test data set is a *particular subset* of the general training data set.

On the other hand, this problem is widely recognized in other domains, such as human learning theory [8] or *natural language processing* [9]. In those domains, research is done in the line of finding the context of the application given a wide training set. In particular, a general language training corpus has to be changed for domain-specific tasks. This task is called *adaptation*.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 11–18, 2005.

On the image domain, the work of Kumar et al. [10] use an EM based refinement of a generative model to infer the particularities of the new data, in what they call *specification*.

Here, the *particularization* problem refers to all those problems in which we can assume the test set to be intrinsically highly correlated and that it is not represented by the overall training set. A common descriptor in the examples described earlier is the fact that the intra-variability of the particular data is smaller than the inter-variability of the overall training set. This *a priori* knowledge is the basic premise for the problem we propose and aim to solve.

In order to exploit this knowledge we will use a mixed approach of supervised and unsupervised processes. The supervised process takes into account the general decision rule (the inter-variability of the complete training set), while the unsupervised process tries to uncover structure of the data (the intra-variability of the test set). The strategy used will be to express both processes using the same framework. The supervised rule will be used to deform the classification space, while the unsupervised process gathers data together in the deformed space. This process is called **Supervised Clustering Competition Scheme**, *SCCS* from now on.

2 Supervised Clustering Competition Scheme

2.1 General Framework and Main Goal

We aim to find an integrated framework for supervised and unsupervised classification processes so that both can compete for the classification of a data point. In this sense, the clustering methods will lead the data based on the minimization of a dissimilarity measure or maximization of a similarity measure, and the supervised classification process has to guide the data according to the classification borders. Since both techniques has to compete, we cast both processes in the same minimization framework. Though, the clustering scheme can be casted into it straightforwardly, the same does not happen to the supervised process. This last process must be reformulated. Therefore, in order to blend both processes we use the following equation,

$$L(\mathbf{x}) = h(min_{\mathbf{x}}(\alpha \cdot SF(\mathbf{x}) + (1-\alpha) \cdot UF(\mathbf{x}))) \tag{1}$$

where SF stands for *supervised functional*, a functional the minimums of which are close to the centers of each class and that has an explicit maximum on the border between classes; UF stands for *unsupervised functional*, expressing a dissimilarity measure; α is the mixing parameter; and the function $h(\cdot)$ is a decision scheme that allows the classification of each data sample.

To minimize this process we can use gradient descent,

$$\frac{\partial \mathbf{x}}{\partial t} = -\nabla F(\mathbf{x})$$

The iterative scheme is,

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \triangle t \cdot \nabla F(\mathbf{x})$$

where $\nabla F(\mathbf{x}) = \{\partial F(\mathbf{x})/\partial x_i\}$. Therefore,

$$\frac{\partial \mathbf{x}}{\partial t} = -\alpha \frac{\nabla SF(\mathbf{x})}{\|\nabla SF(\mathbf{x})\|} - (1-\alpha)\frac{\nabla UF(\mathbf{x})}{\|\nabla UF(\mathbf{x})\|} \tag{2}$$

The next step is to define the minimization process that represent the supervised classification and the clustering process.

Similarity Clustering. When considering the unsupervised functional, we are interested in tools for robust clustering related to invariability of the initial state, capability to differentiate among different volumes of different sizes and toleration to noise and outliers [1–3].

In the approach we have chosen [1] a *similarity* measure between two points $S(z_j, x_i)$ is used in a maximization framework. Our goal is to maximize the total similarity measure $J_s(\mathbf{x})$ defined as:

$$J_s(\mathbf{x}) = \sum_{i=1}^{c}\sum_{j=1}^{n} f(S(z_j, x_i))$$

where $f(\cdot)$ is a monotone increasing function, $\mathbf{x}$ represents the centers of each cluster and $\mathbf{z}$ is the original data set (where $\mathbf{z} = \{z_1, \ldots, z_n\}$ and z_i is a D-dimensional data point). As a similarity relation $S(z_j, x_i)$, we use,

$$S(z_j, x_i) = e^{-(\frac{\|z_j - x_i\|^2}{\beta})}$$

where β is a normalization term. Let the monotone increasing function $f(\cdot)$ be,

$$f(\cdot) = (\cdot)^{\gamma} \quad , \quad \gamma > 0$$

Therefore, the complete similarity measure $J_s(\mathbf{x})$ is,

$$J_s(\mathbf{x}) = \sum_{i=1}^{c}\sum_{j=1}^{n} \left(e^{-\frac{\|z_j - x_i\|^2}{\beta}}\right)^{\gamma} \tag{3}$$

The parameter β is superfluous in this scheme and can be defined as the sample variance,

$$\beta = \frac{\sum_{j=1}^{n} \|z_j - \bar{z}\|^2}{n} \quad \text{where} \quad \bar{z} = \frac{\sum_{j=1}^{n} z_j}{n}$$

The parameter γ gains a considerable importance in this scheme since a good γ estimate induces a good clustering result. The process of maximizing the total similarity measure is a way to find the peaks of the objective function $J_s(\mathbf{x})$. It is shown in [1] that the parameter γ is used as a neighboring limiter, as well as a local density approximation. To find γ one can use an exhaustive search of the correlation of the similarity function for each point when changing the parameter. If the correlation value is over a certain threshold, we can consider that the similarity measure represents the different variability and volumes of the data set accurately. The authors set this threshold experimentally to 0.97 but can change according to the application.

The similarity clustering approach uses the same similarity function but, as it is a self-organizing approach, we define the initial data and centers by the unlabelled data points $\mathbf{z}^0 = \mathbf{x}^0$,

$$UF(\mathbf{x}) = J_s(\mathbf{x}) = \sum_{j=1}^{n} \big(e^{-\frac{\|z_j - x_k\|^2}{\beta}}\big)^{\gamma}, \quad k = 1 \ldots n$$

Getting its gradient, we obtain,

$$\nabla UF(\mathbf{x}) = -2\frac{\gamma}{\beta}\sum_{j=1}^{n} \big(e^{-\frac{\|z_j - x_k\|^2}{\beta}}\big)^{\gamma}(z_j - x_k), \quad k = 1 \ldots n \tag{4}$$

2.2 Supervised Classifier Functional

The definition of the supervised classifier functional should be made so that both processes can interact with each other. Therefore, we must reformulate the supervised classifier process as a self-organizing iterative process.

Without loss of generality, we restrict our classifier design to a two class supervised classification process using a Bayesian framework with known class probability density functions. Assuming that we can estimate each class probability density function $f_A(\mathbf{x}|c = A)$ and $f_B(\mathbf{x}|c = B)$, the optimal discriminative rule using a Bayesian approach is given by,

$$h(\mathbf{x}) = \begin{cases} A & f(\mathbf{x}|c = A)P(A) > f(\mathbf{x}|c = B)P(B) \\ B & f(\mathbf{x}|c = A)P(A) \leq f(\mathbf{x}|c = B)P(B) \end{cases}$$

If *a priori* knowledge of the probability appearance is not known, we assume $P(A) = P(B)$.

The manifold we are looking for, must maintain the optimal borderline as a maximum, since we want the minimums to represent each of the classes. It can be easily seen that the following family of functions satisfies the requirement,

$$SF = -(f(\mathbf{x}|c = A) - f(\mathbf{x}|c = B))^{2N}, \qquad \forall N \in \{1..\infty\} \tag{5}$$

In order to obtain a feasible closed form of the functional we can restrict the density estimation process to a Gaussian mixture model,

$$SF(\mathbf{x}) = \mathbf{f_K}(\mathbf{x}) = \sum_{\mathbf{i=1}}^{\mathbf{M_k}} \pi_{\mathbf{i}}\mathbf{g_i}(\mathbf{x}, \theta_{\mathbf{i}})$$

where M_k is the model order and $g_i(\mathbf{x}, \theta_i)$ is the multidimensional gaussian function,

$$g_i(\mathbf{x}, \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|}e^{-\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i)}$$

where $\theta_i = \{\mu_i, \Sigma_i\}$ are the parameters for the gaussian, the covariance matrix and the mean, and d is the dimensionality of the data.

2.3 The Procedure

Summarizing the overall procedure, we begin the process with three data sets, the labelled data, the unlabelled data and the test set.

Labelled data is used to create the model for the supervised functional as well as used for the final decision step. Unlabelled data feeds the competition scheme and it is the data that will be labelled according to the balance of the supervised and unsupervised processes. The final algorithm is as follows:

Step 1: Determine a supervised set L of data among labelled data and a set U of unlabelled data.
Step 2: Estimate $SF(\mathbf{x})$ $x \in L$ from (5).
Step 3: Apply the CCA[a] step to set the parameter γ of $UF(x)$ $x \in U$.
Step 4: Feed the competition scheme by evolving the clusters (that come from unlabelled data U) according to (2) and let them converge.

[a] Please refer to [1] for further information on this process.

3 Experimental Results

To illustrate the behavior and the advantages of the Supervised Clustering Competition Scheme in a real particular problem, we have performed our tests on data from the MNIST handwritten characters database.

3.1 Experiment Settings

In order to exemplify the behavior of the methodology we have chosen a full supervised classification scenario in which an OCR is used to distinguish between two hand written digits. In order to ensure structure in the data, we aim for the classification of data of a single writer at a time, in the same way as if the OCR was scanning and recognizing the digits in a single sheet of paper.

We have used the MNIST digits database with images of 128 × 128 pixels. The feature extraction process has been a 4×4 zoning procedure on the 64×64 central values of the image. The number of white pixels is counted and stored as a feature. Therefore we have a 16 dimensional feature space. The training set is composed by 100 different hand-written sets of 120 digits each one (The first 100 writers of the HSF7). Each set corresponds to a different writer. The test set is a 500 different hand-written sets of 120 digits each one (The first 500 writers of the HSF6). We center the experiment in a two-class classification process distinguishing number ONE from number TWO. From the set of features we have selected automatically two of the most discriminative ones using discriminant analysis.

Figure 1 shows the two class classification problem. The figure represents the feature space associated to the training set of the digits “two” and “one”. As one can see, the training set is general and contains different instances of the digit “two” with different particularities. However, not always a general purpose

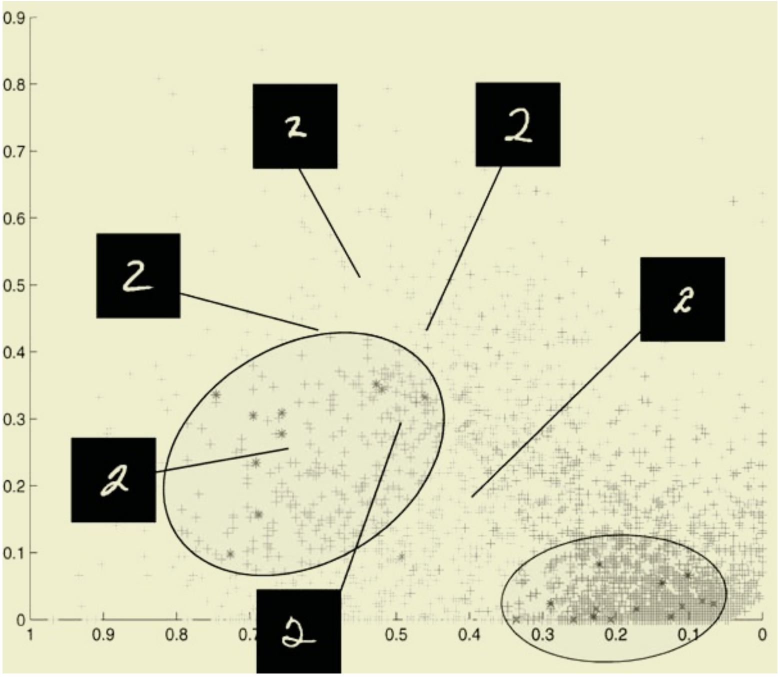


Fig. 1. Feature space representation of the training set. The asterisks represents the test set. The shadow areas represents a possible partition of the training set that solves the two class problem.

classifier is needed. If our particular instance of the problem is a subset of the general set, we can focus on extracting some information of the test data, so we can constrain the training space. The asterisks are the test data for the digits "one" and "two" written by the same writer. The shadowed areas represent possible subsets of the training data which characterize our test data. We can see that this subproblem has much more structure than the original one and that the intra-variability of the particular set is lower than the variability for all writers. In this scenario we can use the SCCS.

Figure 2 shows a representative set of hand-written digits. Figure 2.a shows the variability of the digit "two" written by different writers. Figure 2.b shows the variability of the digit "two" written by just one author.

Experiment Results. As a ground-truth unsupervised classifier we have used Expectation-Maximization with K-Means initialization, and a supervised labelling for the cluster centers. This process yields a recognition rate of 89.35%. The supervised classifier related to the whole training set achieves a recognition rate of 87.65%. This supervised classifier is the same we are mimicking with the supervised functional, so that we can compare performances of adding the unsupervised process.

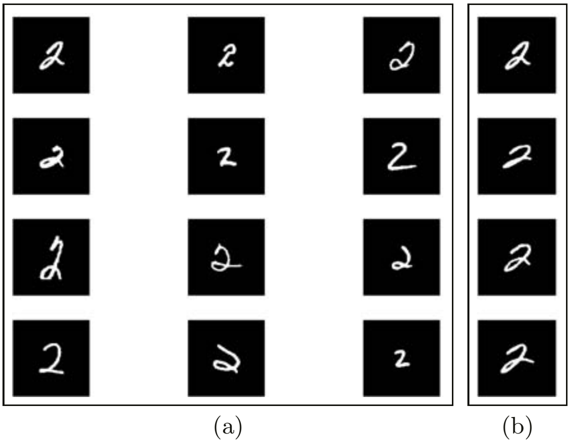


Fig. 2. (a) Set of digit "two" written by different authors. (b) Set of digit "two" written by the same writer.

Table 1. Comparative table for fixed α values.

Mixing value	Recognition Rate	Overall Gain
0	87.73%	0.08%
0.05	87.16%	−0.49%
0.1	87.49%	−0.16%
0.15	87.89%	0.24%
0.2	87.73%	0.08%
0.25	88.14%	0.49%
0.3	89.67%	2.02%
0.35	90.65%	3.00%
0.4	93.07%	5.42%
0.45	92.59%	4.94%
0.5	87.65%	0.00%

Table 1 shows the behavior of the process for different fixed values of the parameter α. The parameter γ is set to adapt automatically. The second column refers to the recognition rate of the overall process. It can be clearly seen that a mixture of supervised and unsupervised improves the performance of the supervised classifier. We must take into account that when $\alpha \geq 0.5$ the process behaves as a pure supervised classifier, thus, we only show the results for α between 0 and 0.5. The third column (Overall Gain) represents the percentage of gain or degradation of the process as the difference between the error of the process and the error achieved using the supervised classifier. Positive values are gain in performance. Negative values are degradation of the process performance. Hence, for $\alpha = 0.1$ the process performs worse than the supervised classifier. This is particularly obvious in the sense that low values of α mean a nearly no contribution of the supervised process, and therefore, the classification is made with the unsupervised process alone. As we can see, the discriminative power of the resulting classifier improves by more than a 5% the classification results obtained by the supervised classifier without the unsupervised part.

4 Conclusions and Future Lines

In this paper we have introduced a new competition technique for pattern recognition that combines supervised and unsupervised approaches. This technique exploits the structure of the data set to be classified and adds this feature to a classical supervised classification process based on generative models.

The method we propose has been used in an emerging problem of *particularization*. The results using real data of MNIST database with 500 writers show that the SCCS improves the performance of the supervised rule alone by more than 5%. This method has been particularized for a multidimensional two class problem, we are now developing the formulation for the basis of the multi-class supervised functional.

References

1. Miin-Shen Yang and Kuo-Lung Wu,A Similarity-Based Robust Clustering Method. IEEE Trans. on PAMI, Vol. 26, No. 4, pp. 434-448, 2004.
2. R.N. Dave and R. Krishnapuram. Robust Clustering Methods: A Unified View. IEEE Trans. Fuzzy Systems, vol. 5, pp. 270-293, 1997.
3. P.J. Huber. Robust Statistics. Wiley, 1981.
4. A. McCallum and K. Nigam. Employing EM and pool-based active learning for text classification. Int. Conf. on Machine Learning, pp. 359-367, 1998.
5. T.J. O'Neill. Normal Discrimination with unclassified observations. J. of American Statistical Assoc, vol. 73, pp. 821-826, 1978.
6. F. Gagliardi and M. Cirelo. Semi-Supervised Learning of Mixture Models. Proc. XXth ICML, 2003.
7. A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. Proc. XIth. Annual Conference on Computational Learning Theory, Madison, WI, 1998.
8. W.J. Clancey.A Tutorial on Situated Learning. In Proc. of the International Conference on Computers and Education, pp. 49-70, 1995.
9. A.R. Coden, S.V. Pakhamov and C.G. Chute. Domain-Specific Language Models and Lexicons for Tagging. Tech. Reprt. RC23195 (W0404-146) IBM Research Division, 2004.
10. S. Kumar, A. Loui and M. Herbert. An observation-constrained generative approach for probabilistic classification of image regions. Image and Vision Computing, vol. 21, pp. 87-97, 2003.

Adaptive Optimization with Constraints: Convergence and Oscillatory Behaviour

Fernando J. Coito[1] and João M. Lemos[2]

[1] Faculdade de Ciências e Tecnologia - Universidade Nova de Lisboa
fjvc@fct.unl.pt
[2] INESC ID-Lisboa
jlml@inesc-id.pt

Abstract. The problem of adaptive minimization of globally unknown functionals under constraints on the independent variable is considered in a stochastic framework. The CAM algorithm for vector problems is proposed. By resorting to the ODE analysis for analysing stochastic algorithms and singular perturbation methods, it is shown that the only possible convergence points are the constrained local minima. Simulation results in 2 dimensions illustrate this result.

1 Introduction

There are engineering optimization problems in which the global form of both the cost functional and the constraints are unknown. In these problems, when the independent variable is settled to a specific value, the corresponding value of the functional can be read and the decision whether the constraints are or are not being violated can be made. The solution amounts to a number of values applied to the plant according to a functional and constraint models, which are adapted from incoming data. Although these extremum seeking methods have already been the subject of early literature in Adaptive Systems – see [3] for a review – they are receiving increasing interest in recent literature.

This kind of problems are solved in [1,6] by using a self-tuning extremum seeker in which the cost functional is locally approximated by a quadratic function and no constraints are assumed in the independent variable. In this work, the above algorithm is extended for incorporating constraints and the use of vector independent variable.

2 Problem Formulation

Let $y(\cdot)$ be a differentiable function of $\Re^2$ in $\Re$. Consider the following problem

<u>*Problem 1*</u> Find $\mathbf{x}^* = \left[x_1^* \quad x_2^*\right]^T$ such that $y\left(\mathbf{x}^*\right)$ is minimum, subject to the set of constraints $\mathbf{g}\left(\mathbf{x}^*\right) \leq \mathbf{0}$ where $\mathbf{g} \in \Re^{\mathrm{n}}$ and $\mathbf{0}$ is the null vector. □

According to the Kuhn-Tucker theorem, Problem 2 is equivalent to the following

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 19–26, 2005.

Problem 2 Define the Lagrangean function $\pounds(\mathbf{x},\boldsymbol{\rho})\overset{\Delta}{=}y(\mathbf{x})+\boldsymbol{\rho}^T\mathbf{g}(\mathbf{x})$.

Find the $\mathbf{x}^*$ minimizing $\pounds(\mathbf{x},\rho^*)$, in which ρ^* is a vector of Lagrange multipliers, satisfying the Kuhn-Tucker complementary condition where $\overset{\bullet}{\times}$ is the *term-by-term* multiplication:

$$\boldsymbol{\rho}^*\overset{\bullet}{\times}\mathbf{g}\left(\mathbf{x}^*\right)=\mathbf{0} \tag{1}$$

□

Hereafter, the following assumption is supposed to hold:

H0. The global form of functions $y(\cdot)$ and $\mathbf{g}(\cdot)$ is unknown and may be possibly time varying. However, for each $\mathbf{x}$, $y(\mathbf{x})$ and $\mathbf{g}(\mathbf{x})$ may be observed, possibly corrupted by observation noise. □

3 The CAM Algorithm

The algorithm that solves Problem 2 must accomplish two tasks: the adjustment of the Lagrange multipliers ρ in order to fulfill the Kuhn-Tucker complementary condition (1) and, once ρ is settled, to adjust $\mathbf{x}(t)$.

3.1 Adjustment of the Lagrange Multiplier

Following the development in [4] ρ is adjusted according to a gradient minimization scheme:

$$\boldsymbol{\rho}(t)=\boldsymbol{\rho}(t-1)+\varepsilon\gamma(t-1)\boldsymbol{\rho}(t-1)\overset{\bullet}{\times}\mathbf{g}(\mathbf{x}(t)) \tag{2}$$

where ε is a vanishing small parameter and $\{\gamma(t)\}$ is a sequence of positive gains [5].

3.2 Adaptive Optimization

H1. It is assumed that, close to $\mathbf{x}^*$, the Lagrangean function $\pounds(\mathbf{x},\rho^*)$ may be approximated by a quadratic function:

$$L(t)\overset{\Delta}{=}\pounds(\mathbf{x}(t),\boldsymbol{\rho})=\pounds^*+\left[\mathbf{x}(t)-\mathbf{x}^*\right]^T\mathbf{A}\left[\mathbf{x}(t)-\mathbf{x}^*\right]+\bar{e}(t) \tag{3}$$

in the sequel it will be assumed $\mathbf{A}=\begin{bmatrix}a_{11} & a_{12}\\ a_{12} & a_{22}\end{bmatrix}$ to be symmetric, which does not affect the problem generality. $\mathbf{A}$, $\pounds^*$ and $\mathbf{x}^*$ are unknown parameters, which depend on the value of ρ; $\bar{e}$ is a residue.

Define the increments:

$$\begin{aligned}&\Delta L(t)\overset{\Delta}{=}L(t)-L(t-1) &&\Delta x_i(t)\overset{\Delta}{=}x_i(t)-x_i(t-1)\quad;\quad i=1,2\\ &\Delta x_i^2(t)\overset{\Delta}{=}x_i^2(t)-x_i^2(t-1)\quad;\quad i=1,2 &&\Delta[x_1x_2](t)\overset{\Delta}{=}x_1(t)\cdot x_2(t)-x_1(t-1)\cdot x_2(t-1)\end{aligned} \tag{4}$$

Then equation (3) may be written as

$$\Delta L(t)=[\theta_1\ \cdots\ \theta_5][\Delta x_1(t)\ \ \Delta x_2(t)\ \ \Delta x_1^2(t)\ \ \Delta x_2^2(t)\ \ \Delta[x_1x_2](t)]^T+e(t) \tag{5}$$

where $\theta_1 = -2a_{11}x_1^* - 2a_{12}x_2^*$ $\quad \theta_3 = a_{11}$ $\quad \theta_3 = 2a_{12}$
$\theta_2 = -2a_{22}x_2^* - 2a_{12}x_1^*$ $\quad \theta_4 = a_{22}$

and $e(t) \overset{\Delta}{=} \bar{e}(t) - \bar{e}(t-1)$ is assumed to be an uncorrelated zero mean stochastic sequence such that all moments exist.

Defining $\boldsymbol{\theta}^* \overset{\Delta}{=} [\theta_1 \quad \cdots \quad \theta_5]$ and

$\boldsymbol{\varphi}(t) \overset{\Delta}{=} [\Delta x_1(t) \quad \Delta x_2(t) \quad \Delta x_1^2(t) \quad \Delta x_2^2(t) \quad \Delta[x_1 x_2](t)]^T$ expression (5) yields

$$\Delta L(t) = \boldsymbol{\theta}^* \boldsymbol{\varphi}(t) + e(t) \tag{6}$$

which constitutes a linear regression model in which $\boldsymbol{\theta}^*$ is the vector of coefficients to estimate and $\boldsymbol{\varphi}$ is the data vector.

The vector $\boldsymbol{\theta}^*$ may be estimated using a recursive least-squares algorithm, and the value of **x** that minimizes $L(\mathbf{x})$ is given by:

$$\begin{bmatrix} x_1^* & x_2^* \end{bmatrix}^T = \begin{bmatrix} \dfrac{2\theta_1\theta_4 - \theta_2\theta_5}{\theta_5^2 - 4\theta_3\theta_4} & \dfrac{2\theta_2\theta_3 - \theta_1\theta_5}{\theta_5^2 - 4\theta_3\theta_4} \end{bmatrix}^T \tag{7}$$

3.3 The CAM Algorithm

Combining both the above procedures results in the following *Constrained Adaptive Minimization* (CAM) algorithm:

1. Apply $\mathbf{x}(t)$ to the system and measure $y(t)$ and $\mathbf{g}(\mathbf{x}(t))$
2. Adjust the Lagrange multiplier vector according to equation (2).
3. Build the Lagrangean function associated with the current Lagrange multiplier vector and the current value $y(t)$.
4. Compute the increments (4).
5. Using a RLS algorithm update the estimates of $\boldsymbol{\theta}$ in the model (6).
6. Update the estimates according to

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix}^T = \begin{bmatrix} \dfrac{2\theta_1\theta_4 - \theta_2\theta_5}{\theta_5^2 - 4\theta_3\theta_4} & \dfrac{2\theta_2\theta_3 - \theta_1\theta_5}{\theta_5^2 - 4\theta_3\theta_4} \end{bmatrix}^T + \boldsymbol{\eta}(t) \tag{8}$$

7. Increment the time and go back to step 1.

4 ODE Analysis

The CAM algorithm is now analyzed using the ODE method for analyzing stochastic algorithms [5] and singular perturbation theory for ordinary differential equations [2].

The algorithm is associated with the following set of differential equations:

$$\frac{d\,\boldsymbol{\rho}(t)}{dt} = \varepsilon \boldsymbol{\rho}(t) \overset{\bullet}{\times} \mathbf{g}(\mathbf{x}(t)) \qquad \frac{d\,\boldsymbol{\theta}(t)}{dt} = R^{-1}\mathbf{f}(\boldsymbol{\theta}, \boldsymbol{\rho}) \tag{9}$$

where $R \overset{\Delta}{=} E[\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^T(t)]$ and $\mathbf{f}(\boldsymbol{\theta}, \boldsymbol{\rho}) \overset{\Delta}{=} E\{\boldsymbol{\varphi}(t)[\Delta L(t) - \boldsymbol{\varphi}^T(t)\boldsymbol{\theta}]\}$.

Define the functions $\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\rho})$ and $\mathbf{H}(\boldsymbol{\theta},\boldsymbol{\rho})$

$$\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\rho})\overset{\Delta}{=}R^{-1}\mathbf{f}(\boldsymbol{\theta},\boldsymbol{\rho}) \qquad \mathbf{H}(\boldsymbol{\theta},\boldsymbol{\rho})\overset{\Delta}{=}\overset{\bullet}{\boldsymbol{\rho}}\times\mathbf{g}(\mathbf{x}(t)) \tag{10}$$

Making use of (10) and changing the time scale by $\tau=\varepsilon t$, equations (9) may then be written in the standard form for singular perturbation analysis:

$$\frac{d\,\boldsymbol{\rho}(\tau)}{d\tau}=\mathbf{H}(\boldsymbol{\theta},\boldsymbol{\rho}) \qquad \varepsilon\frac{d\,\boldsymbol{\theta}(\tau)}{d\tau}=\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\rho}) \tag{11}$$

According to the ODE theory exposed in [5], the only possible convergence points of the CAM algorithm are the equilibrium points of (11), such that the Jacobian matrix has all its eigenvalues in the left complex half-plane:

$$\mathbf{J}=\begin{bmatrix}\dfrac{\partial\mathbf{H}}{\partial\boldsymbol{\rho}} & \dfrac{\partial\mathbf{H}}{\partial\boldsymbol{\theta}}\\ \dfrac{\partial\mathbf{G}}{\partial\boldsymbol{\rho}} & \dfrac{\partial\mathbf{G}}{\partial\boldsymbol{\theta}}\end{bmatrix} \tag{12}$$

H2. The disturbance signal $\boldsymbol{\eta}$ in (8) ensures the persistent excitation requirement, i.e. $E\left[\boldsymbol{\varphi}(\tau)\boldsymbol{\varphi}^T(\tau)\right]$ is full rank. □

H3. The function $\mathbf{G}\left(\overline{\boldsymbol{\theta}},\boldsymbol{\rho}^*\right)$ has isolated real roots □

The equilibrium points of (11) are characterized by one of the following conditions:

A-equilibria

$$\boldsymbol{\rho}=\mathbf{0} \qquad f(\boldsymbol{\theta},\mathbf{0})=\mathbf{0} \tag{13}$$

B-equilibria

$$\mathbf{g}(\mathbf{x})=\mathbf{0} \quad \text{and thus } \boldsymbol{\rho}=\boldsymbol{\rho}^* \qquad f\left(\boldsymbol{\theta},\boldsymbol{\rho}^*\right)=\mathbf{0} \tag{14}$$

4.1 Analysis of the A-Equilibria

If (13) holds the constrained minimum equals the unconstrained minimum. The constrained minimum is therefore interior to the region defined by the set of constraints.

If the persistent excitation requirement holds, as $\frac{\partial\mathbf{H}}{\partial\boldsymbol{\theta}}=\mathbf{0}$, the Jacobian matrix (12) becomes lower triangular and its eigenvalues are the ones of $\frac{\partial\mathbf{H}}{\partial\boldsymbol{\rho}}=[\mathbf{g}(\mathbf{x})]_D$ and $\frac{\partial\mathbf{G}}{\partial\boldsymbol{\theta}}=-\mathbf{I}$, where $\mathbf{I}$ is the diagonal unit matrix and $[\mathbf{g}(\mathbf{x})]_D$ is a diagonal matrix whose elements are the $g_i(\mathbf{x})$. As $\rho_i=0$ which implies $g_i(\mathbf{x})<0$, all the Jacobian eigenvalues have negative real parts. Thus the only possible convergence points are solutions of Problem 1.

4.2 Analysis of the B-Equilibria

If (14) holds the constrained minimum is different from the unconstrained minimum, being located on the boundary of the region defined by $\mathbf{g}(\mathbf{x})\leq\mathbf{0}$. In this case $\frac{\partial\mathbf{H}}{\partial\boldsymbol{\theta}}$ is no longer null. Thus, the Jacobian matrix is not lower triangular, and the analysis from the previous section does not hold.

Making use of the singular perturbation theory (Kokotovic, *et al.*, 1986), assuming that the parameter ε in (2) is vanishing small the two equations in (11) may be seen as the slow and fast subsystems, respectively.

Assume that H3 holds and consider the boundary layer correction $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} - \bar{\boldsymbol{\theta}}$ whose dynamics is

$$\frac{d\tilde{\boldsymbol{\theta}}}{d\tau} = \frac{1}{\varepsilon}\mathbf{G}\left(\tilde{\boldsymbol{\theta}}, \boldsymbol{\rho}^*\right) \tag{15}$$

H4. Assume that $\tilde{\boldsymbol{\theta}}(\tau) = \mathbf{0}$ is an equilibrium point of (15), asymptotically stable, uniformly in $\boldsymbol{\rho}^*$, and that $\boldsymbol{\theta}(0) - \bar{\boldsymbol{\theta}}(0)$ belongs to its domain of attraction. □

Proof of H4: It follows from $\frac{d\tilde{\boldsymbol{\theta}}}{d\tau} = \frac{1}{\varepsilon}R^{-1}E\left\{\boldsymbol{\varphi}(t)\left[\boldsymbol{\varphi}^T(t)\left(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) + e(t)\right]\right\} = -\frac{1}{\varepsilon}R^{-1}R\tilde{\boldsymbol{\theta}} = -\frac{1}{\varepsilon}\tilde{\boldsymbol{\theta}}$ □

H5. The eigenvalues of $\frac{\partial \mathbf{G}}{\partial \boldsymbol{\theta}}$, calculated for ε=0, have strictly negative real part. □

Proof of H5: It results from $\frac{\partial \mathbf{G}}{\partial \boldsymbol{\theta}} = -\frac{\partial}{\partial \boldsymbol{\theta}}R^{-1}E\left\{\boldsymbol{\varphi}(\tau)\boldsymbol{\varphi}^T(\tau)\right\}\tilde{\boldsymbol{\theta}} = -\frac{\partial}{\partial \boldsymbol{\theta}}\tilde{\boldsymbol{\theta}} = -\mathbf{I}$ □

Since these assumptions hold, Tikhonov’s theorem (Kokotovic, *et al.*, 1986) allows to conclude that the only piossible convergence points of the CAM algorithm are the constrained minima of the optimization problem 1.

5 Simulation Results

The ODE analysis characterizes the possible convergence points of the CAM algorithm. Yet, it does not prove that the algorithm will actually converge. In order to exhibit the algorithm convergence features, a number of simulations are presented.

5.1 Example 1

In this example Problem 1 is considered, in which

$$y(\mathbf{x}) = (\mathbf{x} - \mathbf{x_0})^T(\mathbf{x} - \mathbf{x_0}) \qquad \text{where } \mathbf{x_0} = [0.6 \qquad 0.8]^T \tag{16}$$

$$\begin{array}{lll} g_1(\mathbf{x}) = 3 - x_2 e^{x_1/3} \le 0 & g_3(\mathbf{x}) = x_1 - 1 \le 0 & g_5(\mathbf{x}) = x_2 - 3 \le 0 \\ g_2(\mathbf{x}) = -x_1 \le 0 & g_4(\mathbf{x}) = 2 - x_2 \le 0 & \end{array} \tag{17}$$

The identification is performed using RLS with exponential forgetting factor.

Figure 1 presents the evolution of the optimum estimate towards the feasibility region. The constrained minimum is on the frontier of the region. Thus while the Lagrange multipliers related to the inactive constraints go to zero $(\rho_i \to 0)$, those related to active constraints converge to the optimum $\rho_j \to \rho^*$ (figure 1.*c*).

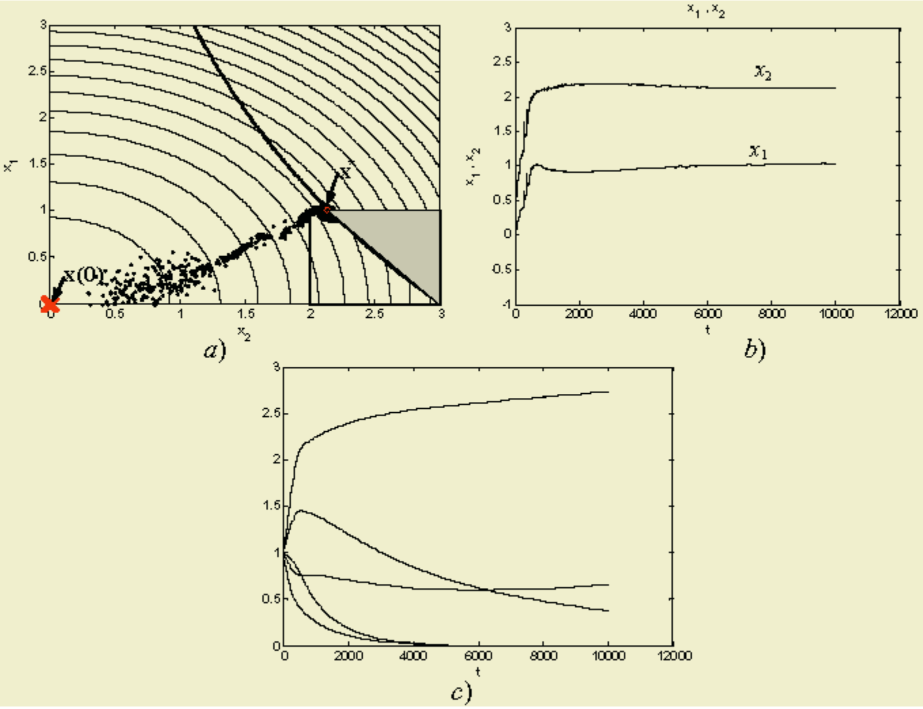


Fig. 1. Adaptive optimum search from example 1. *a*) The gray area is the feasibility region. *b*) x_1 and x_2 time evolution. *c*) Evolution of the Lagrange multipliers

5.2 Example 2: Multiple Local Minima

The ODE analysis presented states that the convergence points are local minima from the constrained optimization problem. Thus, it is interesting to see what occurs when more than one minimum exists. In this example the function to be minimized is given by $y(\mathbf{x}) = 9 + \frac{9}{2}x_1 - 4x_2 + x_1^2 + 2x_2^2 - 2x_1x_2 + x_1^4 - 2x_1^2x_2$ and it is subject to the constraint $g(\mathbf{x}) = 24.25 - \left(x_1^2 + x_2^2\right) \le 0$.

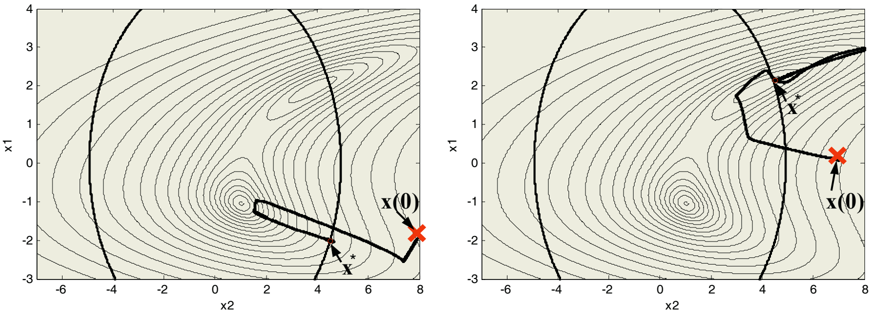
Experiments using the updating scheme from equation (8) have shown that with this scheme assumption H1 and equation (6) would not hold. Thus in the experiment presented the updating scheme (8) was replaced by the following gradient scheme:

$$\mathbf{x}(t+1) = \mathbf{x}(t) - \delta \frac{\partial L}{\partial \mathbf{x}} \Bigg/ \left\| \frac{\partial L}{\partial \mathbf{x}} \right\| \tag{18}$$

Figure 2.*a* presents the algorithm evolution when it starts from the initial point **x**(0)=[-1.9 7.95]^T. It converges towards a local minimum, located at $\mathbf{x}^* = [-2 \quad 4.51]^T$, with a value of the objective function of 19.6.

In figure 2.*b* the algorithm is started from a different initial point, **x**(0)=[0.198 6.95]^T. In this case it converges to another local minimum located at **x***=[2.14 4.51]^T, which corresponds to a value of the objective function of 1.21 (the absolute constrained minimum).

The minimum to which the algorithm converges depends on the initial point **x**(0), and in which domain of attraction it lies.



a) *b*)

Fig. 2. Adaptive optimum search for Example 2. The feasibility region lies outside the bold line. *a*) **x**(0)=[-1.9 7.95]^T *b*) **x**(0)=[0.198 6.95]^T

5.3 Example 3: Improved Performance

In example 2, the adaptation presents a strong transient that strongly violates the feasibility region. This results from the dynamics (2) of the Lagrange multiplier ρ. In order to improve the algorithm performance the gain $\gamma(t)$ is computed according to the following scheme:

$$\gamma(t) = \mathrm{sat}\left\{ \min, \max, \frac{\left[\lambda_\gamma \gamma(t-1) + k_\gamma \, \mathrm{g}(\mathbf{x}(t))\right]}{\mathrm{sat}\{f_\gamma, \infty, \rho(t)\}} \right\} \tag{19}$$

where sat{*min*, *max*, *z*} corresponds to the function that saturates *z* between the values *min* and *max*.

Results of the algorithm performance with this modification are presented in figure 3. The adaptation converges towards a local minimum and does not strongly violate the feasibility region.

The values of λ_γ and f_γ are chosen so that $\gamma(t)$ changes rapidly between its max and min values $(\lambda_\gamma / f_\gamma > 1)$. In the example λ_γ=0.95 and f_γ=0.1 were used. Figure 3.b shows the steady state behaviour. The algorithm doesn't actually converge to a value. Instead it oscillates between the two sides of the constraint border.

6 Conclusion

The problem of adaptive minimization of globally unknown functionals under constraints on the independent variable was addressed in a stochastic framework. The CAM algorithm for vector problems was proposed. By resorting to the ODE analysis for analysing stochastic algorithms and singular perturbation methods, it was shown

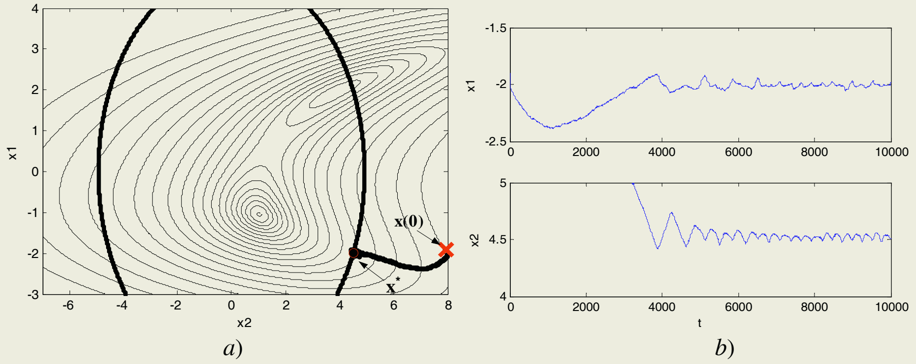


Fig. 3. Adaptive optimum search for Example 3. *a*) The transient does not violate the feasibility region. *b*) A detail of the algorithm steady state

that the only possible convergence points are the constrained local minima. A number of simulation results in 2 dimensions were presented to illustrate this result. Modifications to the original algorithm were introduced to improve performance.

References

1. Bozin, A. S. and M. B. Zarrop (1991). Self-tuning extremum optimizer – Convergence and robustness properties". *Prep. ECC 91*, 672-677.
2. Kokotovic, P., H. K. Khalil and J. O'Reilly (1986). *Singular Perturbation Methods: Analysis and Design*. Academic Press. 1986.
3. Krstič, M. (2000) Performance improvement and limitations in extremum seeking control. *Systems & Control Letters*, **39**, 313-326.
4. Lemos, J. M (1992). Adaptive Optimization with Constraints. *Prep. 4th IFAC Symp. Adaptive Systems in Control and Signal Processeing*, 53-58.
5. Ljung, L. (1977). Analysis of Recursive Stochastic Algorithms. *IEEE Transactions on Automatic Control*, **AC-22**, N°4, 551-575.
6. Wellstead, P. E. and P. G. Scotson (1990). Selftuning extremum control. *IEE Proceedings*, **137**, Pt D. N° 3, 165-175.

Data Characterization for Effective Prototype Selection

Ramón A. Mollineda, J. Salvador Sánchez, and José M. Sotoca

Dept. Llenguatges i Sistemes Informàtics, Universitat Jaume I,
Av. Sos Baynat s/n, E-12071 Castelló de la Plana, Spain
{mollined,sanchez,sotoca}@uji.es

Abstract. The Nearest Neighbor classifier is one of the most popular supervised classification methods. It is very simple, intuitive and accurate in a great variety of real-world applications. Despite its simplicity and effectiveness, practical use of this rule has been historically limited due to its high storage requirements and the computational costs involved, as well as the presence of outliers. In order to overcome these drawbacks, it is possible to employ a suitable prototype selection scheme, as a way of storage and computing time reduction and it usually provides some increase in classification accuracy. Nevertheless, in some practical cases prototype selection may even produce a degradation of the classifier effectiveness. From an empirical point of view, it is still difficult to know a priori when this method will provide an appropriate behavior. The present paper tries to predict how appropriate a prototype selection algorithm will result when applied to a particular problem, by characterizing data with a set of complexity measures.

1 Introduction

One of the most widely studied non-parametric classification approaches corresponds to the k-Nearest Neighbor (k-NN) decision rule [3]. Given a set of n previously labeled instances (training set, TS), the k-NN classifier consists of assigning an input sample to the class most frequently represented among the k closest instances in the TS, according to a certain dissimilarity measure. A particular case of this rule is when $k = 1$, in which an input sample is assigned to the class indicated by its closest neighbor.

The asymptotic classification error of the k-NN rule (i.e., when n grows to infinity) tends to the optimal Bayes error rate as $k \to \infty$ and $k/n \to 0$. Moreover, if $k = 1$, the error is bounded by approximately twice the Bayes error [3]. The optimal behavior of this rule in asymptotic classification performance along with a conceptual and implementational simplicity make it a powerful classification technique capable of dealing with arbitrarily complex problems, provided that there is a large enough TS available.

Nevertheless, this theoretical requirement of large TS size is also the main problem using the 1-NN rule because of the seeming necessity of a lot of memory and computational resources. This is why numerous investigations have been concerned with finding new approaches that are efficient with computations. Within this context, many fast algorithms to search for the NN have been proposed. Alternatively, some prototype selection techniques [1, 4, 6] have been directed to reduce the TS size by selecting only the most relevant instances among all the available ones, or by generating new prototypes in locations accurately defined.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 27–34, 2005.

On the other hand, in many practical situations the theoretical accuracy can hardly be achieved because of certain inherent weaknesses that significantly reduce the effective applicability of k-NN classifiers in real-world domains. For example, the performance of these rules, as with any non-parametric classification approach, is extremely sensitive to data complexity. In particular, class-overlapping, class-density, and incorrectness or imperfections in the TS can affect the behavior of these classifiers. Other prototype selection methods [5, 10, 13, 14] have been devoted to improve the 1-NN classification performance by eliminating outliers (i.e., noisy, atypical and mislabeled instances) from the original TS, and by reducing the possible overlapping between regions from different classes.

Despite the apparent benefits of most prototype selection algorithms, in some domains these techniques might not achieve the expected results due to certain data characteristics. For this reason, it seems interesting to know a priori the conditions under which the application of a prototype selection scheme can become appropriate. A set of data complexity measures [7, 8] are used in this paper to predict when a prototype selection technique leads to an improvement with respect to the plain 1-NN rule.

2 Data Complexity Measures

The behavior of classifiers is strongly dependent on data complexity. Usual theoretical analysis consists of searching accuracy bounds, most of them supported by impractical conditions. Meanwhile, empirical analysis is commonly based on weak comparisons of classifier accuracies on a small number of unexplored data sets. Such studies usually ignore the particular geometrical descriptions of class distributions to explain classification results. Various recent papers [7, 8] have introduced the use of measures to characterize the data complexity and to relate such descriptions to classifier performance.

In [7, 8], authors define some complexity measures for two classes. For our purposes, a generalization of such measures for the n-class problem is accomplished. The ideal goal is to represent classification problems as points in a space defined by a number of measures, where clusters can be related to classification performances. Next paragraphs describe the measures selected for the present study (the same short notation as in the original paper [7] is here used).

Generalized Fisher's Discriminant Ratio (F1). The plain version of this well-known measure computes how separated are two classes according to a specific feature. It compares the difference between class means with the sum of class variances. A possible generalization for C classes, which also considers all feature dimensions, can be stated as follows:

$$F1 = \frac{\sum_{i=1}^{C} n_i \cdot \delta(m, m_i)}{\sum_{i=1}^{C} \sum_{j=1}^{n_i} \delta(x_j^i, m_i)} \tag{1}$$

where n_i denotes the number of samples in class i, δ is a metric, m is the overall mean, m_i is the mean of class i, and x_j^i represents the sample j belonging to class i.

Volume of Overlap Region (F2). The original measure computes, for each feature, the length of the overlap range normalized by the length of the total range in which all values of both classes are distributed. The volume of the overlap region for two classes is the product of normalized lengths of overlapping ranges for all features. Our generalization sums this measure for all pairs of classes, that is,

$$F2 = \sum_{(c_i,c_j)} \prod_{k} \frac{\min\{\max(f_k,c_i),\max(f_k,c_j)\} - \max\{\min(f_k,c_i),\min(f_k,c_j)\}}{\max\{\max(f_k,c_i),\max(f_k,c_j)\} - \min\{\min(f_k,c_i),\min(f_k,c_j)\}} \tag{2}$$

where (c_i, c_j) goes through all pair of classes, k takes feature index values, while $\min(f_k, c_i)$ and $\max(f_k, c_i)$ compute the minimum and maximum values of feature f_k in class c_i, respectively.

Feature Efficiency (F3). In [7], the feature efficiency is defined as the fraction of points that can be separated by a particular feature. For a two-class problem, the original measure takes the maximum feature efficiency. This paper considers the points in the overlap range (instead of those separated points as in the original formulation). The measure value for C classes is the overall fraction of points in some overlap range of any feature for any pair of classes. Obviously, points in more than one range are counted once. This measure does not take into account the joint contribution of features.

Non-parametric Separability of Classes (N2, N3). The first measure (N2) is the ratio of the average distance to intraclass nearest neighbor and the average distance to interclass nearest neighbor. It compares the intraclass dispersion with the interclass separability. Smaller values suggest more discriminant data. The second measure (N3) is simply the estimated error rate of the 1-NN rule by the leaving-one-out scheme.

Density Measure (T2). This measure does not characterize the overlapping level, but contributes to understand the behavior of some classification problems. It describes the density of spatial distributions of samples by computing the average number of instances per dimension.

3 Prototype Selection

Prototype Selection (PS) techniques have been proposed as a way of minimizing the problems related to the k-NN classifier. They consist of selecting an appropriate reduced subset of instances and applying the 1-NN rule using only the selected examples. Two different families of PS methods exist in the literature: editing and condensing algorithms.

Editing [5, 10, 13–15] eliminates erroneous cases from the original set and "cleans" possible overlapping between regions from different classes, what usually leads to significant improvements in performance. Thus the focus of editing is not on reducing the set size, but on defining a high quality TS by removing outliers. Nevertheless, as a by-product these algorithms also obtain some decrease in size and consequently, a reduction of the computational burden of the 1-NN classifier.

Wilson [14] introduced the first editing proposal. Briefly, this consists of using the k-NN rule to estimate the class of each instance in the TS, and removing those whose class label does not agree with that of the majority of its k neighbors. Note that this algorithm tries to eliminate mislabeled instances from the TS as well as close border instances, smoothing the decision boundaries.

On the other hand, condensing [1, 4, 6, 9, 11, 12] aims at selecting a sufficiently small set of training instances that produces approximately the same performance than the 1-NN rule using the whole TS. It is to be noted that many condensing schemes make sense only when the classes are clustered and well-separated, which constitutes the focus of the editing algorithms.

Hart's algorithm [6] is the earliest attempt at minimizing the number of stored instances by retaining only a *consistent* subset of the original TS. A consistent subset, say S, of a set of instances, T, is some subset that correctly classifies every instance in T using the 1-NN rule. Although there are usually many consistent subsets, one generally is interested in the *minimal* consistent subset (i.e., the subset with the minimum number of instances) to minimize the cost of storage and computing time. Unfortunately, Hart's algorithm cannot guarantee that the resulting subset is minimal in size.

4 Experimental Results and Discussion

As already stated in Sect. 1, in some cases PS algorithms may produce an effect different from the one theoretically expected, that is, they may even degrade the performance of the plain 1-NN classifier. A way of characterizing the problems could be by using the data complexity measures introduced in Sect.2. Thus the experiments reported in this paper aim at describing the databases in terms of such measures and analyzing the conditions under which PS methods can perform better than the plain 1-NN rule.

In our experiments, we have included a total number of 17 data sets taken from the UCI Machine Learning Database Repository (`http://www.ics.uci.edu/~mlearn`) and from the ELENA European Project (`http://www.dice.ucl.ac.be/neural-nets/Research/Projects/ELENA/`). The 5-fold cross-validation error estimate method has been employed for each database: 80% of the available instances have been used as the TS and the rest of instances for the test set. The main characteristics of these data sets and their values for the complexity measures previously described are summarized in Table 1.

For the PS methods, we have tested Wilson's editing, Hart's condensing, and the *combining* edited and condensed set. In this latter case, we have firstly applied Wilson's editing to the original TS in order to remove mislabeled instances and smooth the decision boundaries, and then Hart's algorithm has been used over the Wilson's edited set to further reduce the number of training examples. After preprocessing the TS by means of some PS scheme, the 1-NN classifier has been applied to the test set.

Table 2 reports the error rate and the percentage of original training instances retained by each method for each database. Typical settings for Wilson's editing algorithm (i.e., number of neighbors) have been tried and the ones leading to the best performance have been finally included. The databases are sorted by the value of F1. By means of the data complexity measures, we have tried different orderings which could give us an indication of the relation between the complexity of a data set and the particular method applied to it. From all those measures, it seems that F1 is the one that better discriminates between the cases in which an editing has to be firstly applied and those in which one could directly employ the plain 1-NN rule.

Table 1. Experimental data sets: characteristics and complexity measures.

	Classes	Dim	Samples	F1	F2	F3	N2	N3	T2
Cancer	2	9	683	1.315	0.319	0.902	0.220	0.950	76
Clouds	2	2	5000	0.245	0.380	0.877	0.019	0.846	2500
Diabetes	2	8	768	0.032	0.252	0.994	0.839	0.679	96
Gauss	2	2	5000	0.000	0.309	0.960	0.060	0.650	2500
German	2	24	1000	0.026	0.664	0.992	0.794	0.664	42
Glass	6	9	214	0.474	0.013	0.963	0.452	0.734	24
Heart	2	13	270	0.041	0.196	0.985	0.838	0.567	21
Liver	2	6	345	0.017	0.073	0.968	0.853	0.623	58
Phoneme	2	5	5404	0.082	0.271	0.878	0.067	0.912	1081
Satimage	6	36	6435	2.060	0.000	0.883	0.215	0.909	179
Segment	7	19	2310	0.938	0.000	0.583	0.072	0.967	122
Sonar	2	60	208	0.029	0.000	0.947	0.544	0.827	3
Texture	11	40	5500	3.614	0.000	0.726	0.119	0.992	138
Vehicle	4	18	846	0.259	0.169	0.968	0.273	0.653	47
Vowel	11	10	528	0.536	0.482	0.962	0.129	0.991	53
Waveform	3	21	4999	0.410	0.007	0.997	0.769	0.780	238
Wine	3	13	178	2.362	0.000	0.315	0.018	0.770	14

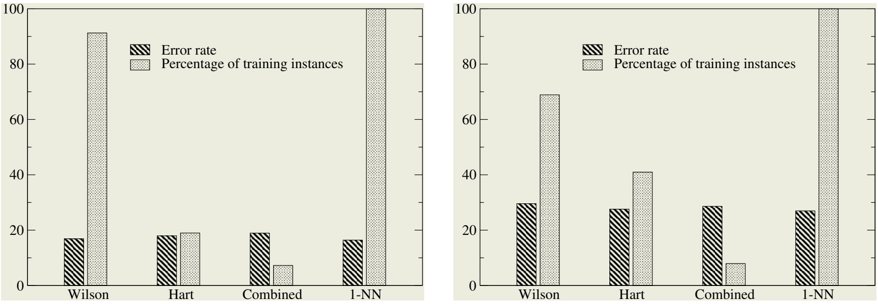
As can be seen in Table 2, Wilson's editing outperforms the 1-NN rule when F1 is under 0.410 (that is, when regions from different classes are strongly overlapped). Consequently, for a particular problem, one could decide to apply an editing to the original TS or directly to employ the plain 1-NN classifier according to the value of F1. For data sets with no (or weak) overlapping (in Table 2, those with F1 > 0.410), the use of an editing can become even harmful in terms of error rate: it seems that editing removes some instances that are defining the decision boundary and therefore, this produces a certain change in the form of such a boundary. Another important result in Table 2 refers to the percentage of training instances given by Hart's condensing: in general, the reductions in TS size for databases with high overlap are lower than those in the case of data sets with weak overlapping.

From the results included in Table 2, it is possible to distinguish between two situations. First, for domains in which the classes are strongly overlapped, one has to employ an editing algorithm in order to obtain a lower error rate (in these cases, benefits in size reduction and classification time are also obtained). Second, for databases with weak overlapping (i.e., F1 is high enough), in which error rate given by the 1-NN rule can be even lower than that achieved with an editing, one should still decide when to apply a PS scheme (reducing time and storage needs) and when to directly use the 1-NN classifier without any preprocessing. In many problems, differences in error rate are not statistically significant (for example, in Satimage database, the error rates for Wilson's editing and 1-NN rule are 16.90% and 16.40%, respectively) and in such cases, savings in memory requirements and classification times can become the key issues for deciding which method to employ.

Table 2. 1-NN error rate and percentage of training instances (in brackets), sorted by F1 (values in italics indicate the lowest error rate for each database).

	F1	Wilson		Hart		Combined		1-NN	
Gauss	0.000	*30.24*	(68.93)	35.86	(54.07)	30.76	(8.08)	35.06	(100.00)
Liver	0.017	*32.18*	(66.59)	37.68	(59.13)	34.17	(17.46)	34.50	(100.00)
German	0.026	*30.60*	(68.10)	38.50	(53.45)	30.49	(10.73)	34.69	(100.00)
Sonar	0.029	43.03	(82.04)	50.40	(34.49)	*40.42*	(17.25)	47.89	(100.00)
Diabetes	0.032	*27.21*	(71.66)	35.29	(51.47)	27.34	(10.78)	32.68	(100.00)
Heart	0.041	*32.61*	(58.06)	42.14	(59.54)	35.20	(13.52)	41.83	(100.00)
Phoneme	0.082	*26.43*	(89.42)	34.07	(21.55)	28.17	(9.28)	29.74	(100.00)
Clouds	0.245	*11.52*	(88.06)	17.28	(27.25)	11.80	(4.07)	15.34	(100.00)
Vehicle	0.259	36.54	(64.15)	36.76	(53.43)	37.36	(18.65)	*35.59*	(100.00)
Waveform	0.410	*18.96*	(82.01)	26.01	(38.96)	21.84	(17.09)	22.04	(100.00)
Glass	0.474	32.37	(70.69)	31.35	(47.01)	32.74	(18.74)	*28.60*	(100.00)
Vowel	0.536	5.23	(96.69)	4.57	(23.40)	8.51	(21.96)	*2.10*	(100.00)
Segment	0.938	5.28	(96.09)	5.88	(13.73)	6.88	(9.90)	*3.72*	(100.00)
Cancer	1.315	*4.25*	(95.54)	6.43	(11.44)	4.39	(3.00)	4.54	(100.00)
Satimage	2.060	16.90	(91.24)	17.94	(18.96)	18.93	(7.23)	*16.40*	(100.00)
Wine	2.362	29.57	(68.89)	27.59	(40.97)	28.60	(7.92)	*26.95*	(100.00)
Texture	3.614	1.22	(98.97)	2.91	(8.01)	2.86	(6.86)	*1.04*	(100.00)

Fig. 1 illustrates the situation just described, comparing the error rate and the percentage of training instances for two databases with a high value of F1. For the Satimage database, differences in error rate are not statistically significant but, in terms of percentage of training instances, the combined approach is clearly the best option: it stores only 7.23% of the original samples and provides an error rate approximately 2% higher than the plain 1-NN rule with the whole TS (100% of instances). Results for the Wine database are similar to those of the Satimage domain, although now differences in error rate are more important when comparing Wilson's editing and 1-NN classifier.



(a) Satimage (b) Wine

Fig. 1. Comparing error rate and percentage of the original instances retained by each method for several databases with high F1.

As a conclusion, for these cases with high F1, one has to decide whether it is more important to achieve the lowest error rate but without any reduction in storage or to attain a moderate error rate with important savings in memory requirements (and also, in classification times).

Despite F1 results in the complexity measure with the highest discrimination power in the specific framework of PS, it is to be noted that other measures can become especially useful for other different tasks. For example, F2 and F3 (conveniently adapted) could be particularly interesting in the case of feature selection because they could be used as objective functions to pick subsets of relevant features. On the hand, other measures constitute a complement in the analysis of certain problems. In this sense, T2 can help to understand why the plain 1-NN classifier does not perform well in problems with weak overlapping. For example, the 1-NN error rate in Wine database, which corresponds to a problem with almost no overlapping (F1 = 2.362), is high enough (26.95%); this can be explained by the fact that there exists a very small number of training instances per dimension (T2 = 14).

5 Concluding Remarks and Further Extensions

The primary goal of this paper has been to analyze the relation between data complexity and efficiency for the 1-NN classification. More specifically, we have investigated on the utility of a set of complexity measures as a tool to predict whether or not the application of some PS algorithm results appropriate in a particular problem.

After testing different data complexity measures, from the experiments carried out over 17 databases, it seems that F1 can become especially useful to distinguish between the situations in which a PS technique is clearly needed and those in which a more extensive study has to be considered. While in the former case the PS approach achieves the lowest error rate and some savings in memory storage, for the later it is not clear the significance of gains in error rate and therefore, other measures should be employed because even the application of a method with a higher error rate could be justified according to other benefits in computational requirements.

It is worth noting that for those situations in which PS degrades the 1-NN accuracy, one could still reduce the (high) computing time associated to the plain 1-NN rule by means of *fast search* algorithms [2]. However, it is known that fast search algorithms can lessen the number of computations during classification but they still maintain the memory requirements.

Future work is mainly addressed to extend the data complexity measures employed in the same framework of the present paper, trying to better characterize the conditions for an appropriate use of PS techniques. A larger number of PS algorithms, both from selection and abstraction perspectives, has also to be tested in order to understand the relation between data complexity and performance of the 1-NN classifier. Finally, a more exhaustive study will help to categorize the use of several complexity measures for different pattern recognition tasks.

Acknowledgments

This has work been supported in part by grants TIC2003-08496 from the Spanish CICYT, GV04A/705 from Generalitat Valenciana, and P1-1B2004-08 from Fundació Caixa Castelló-Bancaixa.

References

1. Chang, C.-L.: Finding prototypes for nearest neighbor classifiers, IEEE Trans. on Computers 23 (1974) 1179-1184.
2. Chavez, E., Navarro, G., Baeza-Yates, R.A., Marroquin, J.L.: Searching in metric spaces, ACM Computing Surveys 33 (2001) 273-321.
3. Cover, T.M., Hart, P.E.: Nearest neighbor pattern classification, IEEE Trans. on Information Theory 13 (1967) 21-27.
4. Dasarathy, B.V.: Minimal consistent subset (MCS) identification for optimal nearest neighbor decision systems design, IEEE Trans. on Systems, Man, and Cybernetics 24 (1994) 511-517.
5. Devijver, P.A., Kittler, J.: Pattern Recognition: A Statistical Approach, Prentice Hall, Englewood Cliffs, NJ (1982).
6. Hart, P.E.: The condensed nearest neighbor rule, IEEE Trans. on Information Theory 14 (1968) 515-516.
7. Ho, T.-K., Basu, M.: Complexity measures of supervised classification problems, IEEE Trans. on Pattern Analysis and Machine Intelligence 24 (2002) 289-300.
8. Bernardo, E., Ho, T.-K.: On classifier domain of competence, In: Proc. 17th. Int. Conf. on Pattern Recognition 1, Cambridge, UK (2004) 136-139.
9. Kim, S.-W., Oommen, B.J.: Enhancing prototype reduction schemes with LVQ3-type algorithms, Pattern Recognition 36 (2003) 1083-1093.
10. Kuncheva, L.I.: Editing for the k-nearest neighbors rule by a genetic algorithm, Pattern Recognition Letters 16 (1995) 809-814.
11. Mollineda, R.A., Ferri, F.J., Vidal, E.: An efficient prototype merging strategy for the condensed 1-NN rule through class-conditional hierarchical clustering, Pattern Recognition 35 (2002) 2771-2782.
12. Ritter, G.L., Woodruff, H.B., Lowry, S.R., Isenhour, T.L.: An algorithm for a selective nearest neighbour decision rule, IEEE Trans. on Information Theory 21 (1975) 665-669.
13. Tomek, I.: An experiment with the edited nearest neighbor rule, IEEE Trans. on Systems, Man and Cybernetics 6 (1976) 448-452.
14. Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data sets, IEEE Trans. on Systems, Man and Cybernetics 2 (1972) 408-421.
15. Wilson, D.R., Martinez, T.R.: Reduction techniques for instance-based learning algorithms, Machine Learning 38 (2000) 257-286.

A Stochastic Approach to Wilson's Editing Algorithm

Fernando Vázquez[1], J. Salvador Sánchez[2], and Filiberto Pla[2]

[1] Dept de Ciencia de la Computación, Universidad de Oriente, Av. Patricio Lumunba s/n,
Santiago de Cuba, CP 90100, Cuba
fvazquez@csd.uo.edu.cu
[2] Dept. Lenguajes y Sistemas Informáticos, Universitat Jaume I, 12071 Castellón, Spain
{sanchez,pla}@uji.es

Abstract. Two extensions of the original Wilson's editing method are introduced in this paper. These new algorithms are based on estimating probabilities from the *k*-nearest neighbor patterns of an instance, in order to obtain more compact edited sets while maintaining the classification rate. Several experiments with synthetic and real data sets are carried out to illustrate the behavior of the algorithms proposed here and compare their performance with that of other traditional techniques.

1 Introduction

Among non-parametric statistical classifiers, the approaches based on neighborhood criteria have some interesting properties with respect to other non-parametric methods. The most immediate advantage makes reference to their simplicity, that is, the classification of a new pattern in the feature space is based on the local distribution of patterns in the training set that surround the targeted point.

The Nearest Neighbor (NN) rule [1] is one of the most extensively studied algorithms within the non-parametric classification techniques. Given a set of previously labeled prototypes (a training set, TS), this rule assigns a sample to the same class as the closest prototype in the set, according to a measure of similarity in the feature space. Another extended algorithm is the *k* nearest neighbors rule (*k*-NN), in which a new pattern is assigned to the class resulting from the majority voting of its *k* closest neighbors. Obviously, the *k*-NN rule becomes the NN rule for *k*=1.

In order to achieve an appropriate convergence of the *k*-NN rule, it is well known its asymptotic behavior with respect to the Bayes classifier for very large TS. On the other hand, the larger the TS, the more computational cost is needed, becoming unaffordable for large data sets.

Prototype Selection (PS) techniques for the *k*-NN rule are aimed at selecting prototypes from the original TS to improve and simplify the application of the NN rule. Within the PS techniques, we can differentiate two main approaches. A first category of techniques try to eliminate from the TS prototypes erroneously labeled, commonly outliers, and at the same time, to "clean" the possible overlapping between regions of different classes. These techniques are referred in the literature to as Editing, and the resulting classification rule is known as Edited NN rule [2].

A second group of PS techniques are aimed at selecting a certain subgroup of prototypes that behaves, employing the 1-NN rule, in a similar way to the one obtained

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 35–42, 2005.

by using the totality of the TS. This group of techniques are the so called Condensing algorithms and its corresponding Condensed NN rule [2].

The application of editing procedures are interesting not only as a tool to reduce the classification error associated to NN rules, but also to carry out any later process that could benefit from a TS with simpler decision borders and reduced presence of outliers in the distributions [5], [7]. The Wilson's editing algorithm [6] constitutes the first formal proposal in these PS techniques, which is still widely used because of its effectiveness and simplicity. The present paper presents a new classification rule based on the distances from a sample to its k-nearest neighbor prototypes. Using this likelihood rule, we present two modifications of Wilson's editing.

2 Editing Algorithms

The common idea to most editing algorithms consists of discarding prototypes that are placed in a local region corresponding to a class different from its [5]. As we will see later, basically the only thing that varies among the various editing algorithms is how they estimate the probability that a sample belongs to a certain class.

All the algorithms employed in this work are based on the k-NN classifier. Thus the k-NN rule can be formally expressed as follows. Let $\{X, \theta\}=\{(x_1,\theta_1), (x_2,\theta_2), \ldots, (x_N,\theta_N)\}$ be a TS with N prototypes from M possible classes, and let $P_j = \{P_{j,i} \, / \, i = 1,2,\ldots, N_j\}$ be the set of prototypes from X belonging to class j. The neighborhood $N_k(\mathbf{x})$ of a sample $\mathbf{x}$ can be defined as the set of prototypes:

$$N_k(\mathbf{x}) \subseteq P \, ; \quad |N_k(\mathbf{x})| = k$$

$$\forall\, p \in N_k(\mathbf{x}) \, , \, q \in P - N_k(\mathbf{x}) \Rightarrow \mathrm{d}(p, \mathbf{x}) \le \mathrm{d}(q, \mathbf{x}); \text{ where } \; P = \bigcup_{i=1}^{M} P_i$$

If we now define a new distance between a point and a set of prototypes such as

$$\mathrm{d}_k (\mathbf{x}, P_i) = k - |N_k(\mathbf{x}) \cap P_i|$$

then the k-NN classification rule can be defined as:

$$\delta_{k\text{-NN}}(\mathrm{x}) = \Theta_{\mathrm{i}} \iff \mathrm{d}(\mathrm{x}, P_i) = \min_{\mathrm{j}=1,2,\ldots,\mathrm{M}} \mathrm{d}_k (\mathrm{x}, P_j)$$

Wilson's Editing

Wilson's editing relies on the idea that, if a prototype is erroneously classified using the k-NN, it has to be eliminated from the TS. Thus, all the prototypes in the TS are used to determine the k nearest neighbors, except the one being considered, that is, it uses the leaving-one-out error estimate. Thus, the Wilson's editing algorithm [6] can be expressed as follows:

```
Initialization: $S \leftarrow X$
*For each prototype* $x_i \in X$ do
        Search for the $k$-nearest neighbors of $x_i$ inside $X - \{x_i\}$
        If $\delta_{k\text{-NN}}(x_i) \ne \theta_i$ then $S \leftarrow S - \{x_i\}$.
```

This algorithm provides a set of prototypes organized in relatively compact and homogenous groups. However, for reduced data sets, it turns out incorrect considering that the estimation made on each prototype is statistically independent, which is the basis for a correct interpretation of the asymptotic behavior of the NN rule.

Holdout Editing

With the aim of avoiding such restrictions, a new editing algorithm was proposed based on Wilson's scheme, but changing the error estimate method. This algorithm, called Holdout editing [2], consists of partitioning the TS in *m* not overlapped blocks, making an estimation for each block *j*, using the block ((*j*+1) module *m*) to design the sort key. This procedure allows to consider statistical independence, whenever $m > 2$.

```
Make a random partition of X in $m$ blocs, $T_1, ..., T_m$
For each block $T_j$ (j = 1 , . . . , $m$):
    For each $x_i \in T_j$
      Search for the $k$ nearest neighbors of $x_i$ in $T_{((j+1) \bmod m)}$
      If $\delta_{k\text{-NN}}(x_i) \neq \theta_i$ then $X \leftarrow X - \{x_i\}$
```

Multiedit

The scheme based on partitions allows the possibility of repeating the editing process a certain number of times, say *f* [2]. In this case, the corresponding algorithm is called Multiedit and consists of repeating the Holdout editing but using the 1-NN rule.

```
1. Initialization : $t \leftarrow 0$
2. Repeat until in the last $t$ iterations ($t > f$) do not take place any prototype
    elimination from the set X.
    2.1 Assign to S the result of applying Holdout editing on X using the
        NN rule.
    2.2 If no new elimination has taken place in 2.1, that is, ($|X| = |S|$),
        then $t \leftarrow t+1$ and go to step 2.
    2.3 else, assign to X the content of S and make $t \leftarrow 0$.
```

For sufficiently large sets, the main advantage of the iterative version is that its behavior is significantly better because of the fact it does not have a dependency on parameter *k*, opposite to the previous algorithm.

The behavior of the editing approaches based on partition gets worse as the size of the TS decreases. This degradation of the effectiveness becomes more significant when increasing the number of blocks by partition [3]. In fact, for relatively small sets, Wilson's editing works considerably better than the Multiedit algorithm.

3 Editing by Estimating Conditional Class Probabilities

For all methods described in previous section, the elimination rule in the editing process is based on the *k*-NN rule. In the editing rules here proposed, the majority voting scheme of the *k*-NN rule is substituted by an estimation of the probability of sample to belong to a certain class.

The rationale of this approach is aimed at using a classification rule based on local information of an instance, like the *k*-NN, but considering the form of the underlying probability distribution in the neighborhood of a point. In order to estimate the values of the underlying distributions, we can use the distance between the sample and the prototypes. Given a sample, the closer a prototype, the more likely this sample belongs to the same class as the one of such a prototype.

Therefore, let us define the probability $P_i(\mathbf{x})$ that a sample **x** belongs to a class *i* as:

$$P_i(x) = \sum_{j=1}^{k} p_i^j \frac{1}{(1+d(x,x^j))}$$

where p_i^j denotes the probability that the *k*-nearest neighbor x^j belongs to class *i.* Initially, the values of p_i^j for each prototype are set to 1 for its class label assigned in the TS, and 0 otherwise. These values could change in case an iterative process is used, but this is not the case in the approach we are presenting here.

The meaning of the above expression states that the probability that a sample **x** belongs to a class *i* is the weighted average of the probabilities that its *k*-nearest neighbors belong to that class. The weight is inversely proportional to the distance from the sample to the corresponding *k*-nearest neighbor. After normalizing,

$$p_i(x) = P_i(x) / \sum_{j=1}^{M} P_j(x)$$

the class *i* assigned to a sample **x** is estimated by the decision rule

$$\delta_{\text{k-prob}}(\mathrm{x}) = i \; ; \quad i \,/\, p_i(x) = \arg\max_j(p_j(x))$$

Using this rule, we propose the editing algorithms described below applying a Wilson's scheme, that is, if the class assigned by the above decision rule does not coincide with the class label of the sample, this sample will be discarded. As we will show in the experiments, the use of the rule just introduced, instead of the *k*-NN rule, makes the editing procedure to produce a TS with a good trade-off between TS size and classification accuracy, because of the fact that such a decision rule estimates in a more accurate way the values of the underlying probability distributions of the different classes, estimating locally these values from the *k*-nearest neighbor samples.

Editing Algorithm Estimating Class Probabilities (WilsonProb)

```
1 Initialization: $S \leftarrow X$
2 For each prototype $x \in X$ do
        2.1 Search for the $k$ nearest neighbors of $x$ inside $X-\{x\}$
        2.2 If $\delta_{k\text{-prob}}(x) \neq \theta$, then $S \leftarrow S - \{x\}$, $\theta$ denotes the class of the object $x$.
```

Editing Algorithm Estimating Class Probabilities and Threshold (WilsonTh)

A variation of the previous algorithm consists of introducing a threshold, $0<\mu<1$, in the classification rule, with the aim of eliminating those instances whose probability to belong to the class assigned by the rule is not significant. Correspondingly, we are removing samples from the TS that are in the decision borders, where the class conditional probabilities overlap and are confusing, in order to obtain edited sets whose instances have a high probability of belonging to the class assigned in the TS.

```
1 Initialization: S $\leftarrow$X
2 For each prototype $x \in$ X do
    2.1 Search for the $k$ nearest neighbors of $x$ inside X – $\{x\}$
    2.2 If $\delta_{k\text{-prob}}(x) \neq \theta$ or $p_j \leq \mu$, do S $\leftarrow$ S – $\{x\}$, $p_j$ is the maximum of all
        the probabilities of the object $x$ to belong to a class.
```

4 Experimental Results and Discussion

In this section, the behavior of the editing algorithms just introduced is analyzed using 14 real and synthetic databases taken from the UCI Machine Learning Database Repository [4]. The main characteristics of these data sets are summarized in Table 1.

Table 1. A brief summary of the experimental databases

	No. classes	No. features	No. instances
Cancer	2	9	683
Liver	2	6	345
Heart	2	13	270
Wine	3	13	178
Australian	2	42	690
Balance	3	4	625
Diabetes	2	8	786
German	2	24	1002
Glass	6	9	214
Ionosphere	2	34	352
Phoneme	2	5	5404
Satimage	6	36	6453
Texture	11	40	5500
Vehicle	4	18	846

The experiments consist of applying the 1-NN rule to each of the test sets, where the training portion has been preprocessed by means of different editing techniques. In particular, apart from the schemes here proposed, Wilson's editing, the Holdout method and the Multiedit algorithm have been also included in this comparative study. The 5-fold cross-validation method (80% of the original instances have been used as the TS and 20% for test purposes) has been here employed to estimate the overall classification accuracy and size reduction rates.

Table 2 reports the experimental results (classification accuracy and size reduction) yielded by the different algorithms over the 14 databases. These results have been averaged over the five partitions. Bold figures indicate the bests methods in terms of classification accuracy for each database. Italics indicates the highest size reduction. Note that typical settings for the algorithms used in the present study have been tried and the ones leading to the best performance have been finally included in Table 2. In the case of WilsonTh, we provide the results obtained from using different values of the threshold parameter. The results corresponding to the plain NN classifier over the original TS have been also included for comparison purposes.

Table 2. Classification accuracy (acc) and size reduction rate (size) using different editings

		NN	Wils.	Hold.	Mult.	WProb	WilsonTh		
							0.6	0.7	0.8
Cancer	acc	95.60	96.19	96.63	96.63	96.34	96.48	96.63	**96.78**
	size		3.44	4.28	7.43	3.36	4.09	5.49	*7.68*
Liver	acc	65.79	**70.70**	70.40	59.49	68.67	68.97	69.55	68.95
	size		32.89	37.10	75.79	27.89	45.94	61.37	*67.82*
Glass	acc	**71.40**	67.62	66.03	58.63	66.16	63.97	62.29	62.31
	size		28.50	46.14	*61.21*	36.68	20.32	50.58	58.17
Heart	acc	58.16	67.00	**67.34**	66.64	66.26	65.17	65.12	64.78
	size		34.44	38.70	*69.25*	28.51	40.09	53.61	65.09
Vehicle	acc	**64.41**	60.26	63.22	52.81	62.16	61.32	61.08	59.67
	size		36.08	39.83	*66.66*	20.41	43.17	46.01	58.86
Wine	acc	**73.04**	70.90	75.24	72.42	69.69	69.74	69.20	69.20
	size		34.97	30.75	*45.50*	14.60	33.28	35.67	41.43
Ionosphere	acc	**83.46**	82.02	82.31	69.58	81.74	81.74	80.89	80.64
	size		16.66	14.52	*34.11*	18.01	18.01	24.21	25.21
Texture	acc	**98.96**	98.63	98.56	94.62	98.74	98.49	98.29	98.32
	size		1.34	3.69	*15.31*	1.01	1.50	3.17	3.06
Balance	acc	79.20	85.11	85.62	86.41	84.96	86.73	88.50	**89.13**
	size		14.80	14.52	37.04	10.76	24.40	32.08	*38.40*
Australian	acc	65.67	69.27	**70.72**	68.99	69.56	69.70	68.39	68.54
	size		31.88	36.88	*59.52*	25.90	37.02	50.76	57.53
German	acc	64.81	70.40	**72.00**	70.00	70.70	71.10	70.50	70.50
	size		30.50	32.27	54.72	26.90	39.62	52.72	*60.00*
Phoneme	acc	70.26	73.53	74.29	**75.35**	73.42	73.44	74.02	73.99
	size		10.56	16.07	*37.43*	11.98	17.26	24.36	29.15
Satimage	acc	**83.62**	83.29	83.32	82.35	83.09	83.18	83.24	83.50
	size		9.43	10.19	*24.51*	9.25	15.61	19.22	23.90
Diabetes	acc	67.32	73.70	73.69	71.09	74.35	74.60	74.48	**74.74**
	size		26.36	44.40	*55.76*	21.09	37.33	45.47	54.91

The first significant result from this empirical analysis is that the editing algorithms here proposed obtain similar classification accuracy to that of other classical methods. It is especially remarkable the fact that no editing outperforms the plain NN classifier in 6 out of 14 databases, although differences in such cases are not statistically significant. Focusing on these results, it seems rather difficult to draw any conclusion because of the little significant differences among them in terms of accuracy.

Examining the other factor of interest in Table 2, that is, the size reduction, the results show that both Multiedit and the proposed WilsonTh achieve the highest rates in all cases, consequently giving the most important decrease in computational loads in the classification phase. Although Multiedit achieves the highest set size reduction rate almost in all databases (10 out of 14), differences with respect to WilsonTh are only marginal. A final remark from the experiments, and perhaps the most important one, refers to the fact of comparing both classification accuracy and reduction rate simultaneously, WilsonTh outperforms Multiedit in most cases. In other words, the proposed WilsonTh algorithm obtains a better trade-off between accuracy and reduction than that given by Multiedit (or any other editing).

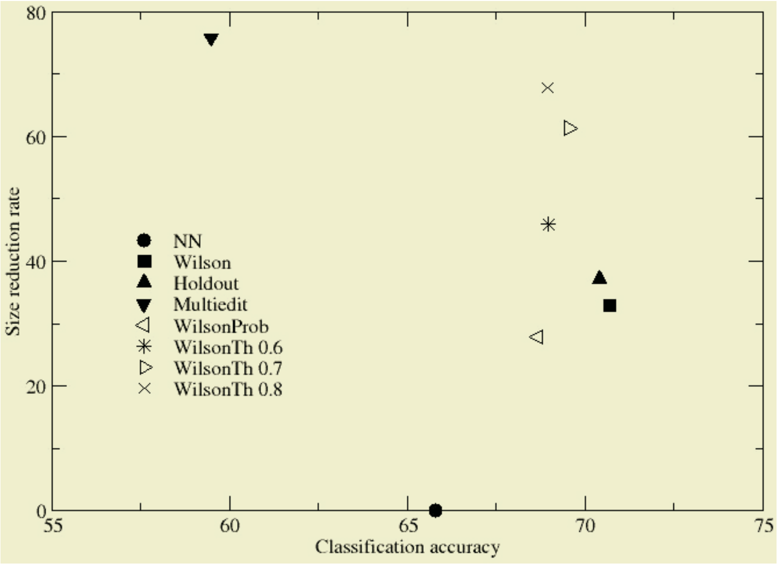


Fig. 1. Comparing classification accuracy and set size reduction for different editing methods over the Liver database

In order to assess the performance relative to these two competing goals simultaneously, Fig. 1 illustrates the behaviour of the editing techniques in terms of both classification accuracy and set size reduction over the Liver database. As can be observed, Multiedit algorithm yields the highest reduction rate, but it produces a very poor classification accuracy. On the contrary, Wilson's editing obtains the highest classification accuracy, but it retains too many training instances. Finally, WilsonTh schemes (0.7 and 0.8) provide a suitable trade-off between both issues: high enough classification accuracy and reduction rate.

5 Concluding Remarks

When using a NN classifier, the presence of mislabeled prototypes can strongly degrade the corresponding classification accuracy. Many models for identifying and removing outliers have been devised. In this paper, we propose two editing algorithms that consider the probabilities of an instance to belong to a class.

A series of experiments over 14 data sets has been carried out in order to evaluate the performance of those new editing methods and compare them with other traditional techniques. From the experiments carried out, it is to be noted that the two stochastic approaches to Wilson's editing attain a suitable trade-off between TS size and classification accuracy.

These editing methods are currently being applied in research works about ongoing learning, where throughout these processes, it is necessary to eliminate erroneously classified instances in the TS, with the objective of improving the classifier, acquiring experience from new unlabeled samples to be incorporated in the TS.

Acknowledgements

This work has been partially supported by projects IST-2001-37306 from EU, TIC2003-08496 from the Spanish CICYT, GV04A/705 from Generalitat Valenciana, and P1-1B2004-08 from Fundació Caixa Castelló-Bancaixa...

References

1. Dasarathy, B. V.: Nearest Neighbor Norms: NN Pattern Classification Techniques. IEEE Computer Society Press, Los Alamos, CA (1991)
2. Devijver, P.A., Kittler, J.: Pattern Recognition: A Statistical Approach. Prentice Hall, Englewood Cliffs, NJ (1982).
3. Ferri, F.J., Albert, J.V., Vidal, E.: Consideration about sample-size sensitivity of a family of edited nearest-neighbor rules. IEEE Trans. on Systems, Man, and Cybernetics-Part B: Cybernetics 29 (1999) 667-672.
4. Merz, C.J., Murphy., Murphy, P.M.: UCI Repository of Machine Learning Database. Department of Information and Computer Science, University of California, Irvine, CA (1998).
5. Sánchez, J.S., Barandela, R., Marqués, A.I., Alejo, R., Badenas, J. : Analysis of new techniques to obtain quality training sets. Pattern Recognition Letters 24 (2003) 1015-1022.
6. Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data sets. IEEE Trans. on Systems, Man and Cybernetics 2 (1972) 408-421.
7. Wilson, D.R., Martinez, T.R.: Reduction techniques for instance-based learning algorithms. Machine Learning 38 (2000) 257-286.

Parallel Perceptrons, Activation Margins and Imbalanced Training Set Pruning

Iván Cantador and José R. Dorronsoro[*]

Dpto. de Ingeniería Informática and Instituto de Ingeniería del Conocimiento,
Universidad Autónoma de Madrid, 28049 Madrid, Spain

Abstract. A natural way to deal with training samples in imbalanced class problems is to prune them removing redundant patterns, easy to classify and probably over represented, and label noisy patterns that belonging to one class are labelled as members of another. This allows classifier construction to focus on borderline patterns, likely to be the most informative ones. To appropriately define the above subsets, in this work we will use as base classifiers the so–called parallel perceptrons, a novel approach to committee machine training that allows, among other things, to naturally define margins for hidden unit activations. We shall use these margins to define the above pattern types and to iteratively perform subsample selections in an initial training set that enhance classification accuracy and allow for a balanced classifier performance even when class sizes are greatly different.

1 Introduction

Most real world classification problems involve imbalanced samples, that is, samples where the number of patterns from one class (that we term the positive samples) is much smaller than that from others. There are many examples of this situation [4, 5], as well as a large literature on this topic, for which many techniques have been applied. Basic examples are ROC curves [8, 12] or the alteration of the sample class distribution, either by oversampling the minority class [2], undersampling the majority class [7] or doing this on both [3]. Moreover, sampling techniques are also in the core of the more sophisticated methods that arise from the well known boosting paradigm [6].

In this work we shall propose a new procedure for training set reduction based on the concept of margin that arises naturally in parallel perceptron (PP) training introduced by Auer et al. in [1]. Parallel perceptrons have the same structure of the well known committee machines [10], that is, they are made up of an odd number of standard perceptrons P_i with ± 1 outputs, and the machine's one dimensional output is simply the sum of these perceptrons' outputs (that is, the overall perceptron vote count). They are thus well suited for 2–class discrimination problems, but it is shown in [1] that they can also be used in regression problems, as they have indeed a universal approximation property.

[*] With partial support of Spain's CICyT, TIC 01–572, TIN2004–07676.

Another contribution of [1] is to give a general and effective training procedure for PPs. A key part of this training procedure is a margin based output stabilization technique that tries to augment the distance of the activation of a perceptron from its decision hyperplane, so that small random changes on an input pattern do not cause its being assigned to another class. Although these margins are not defined on the one dimensional output of a PP and they have to be considered independently for each perceptron, they do provide a way to measure the relevance of individual patterns with respect the overall training set and to establish a pattern selection strategy.

We shall briefly describe in section 2 the training of PPs, as well as their handling of margins, while in section 3 we will describe the overall training set selection procedure and shall also see how margins can be used to discard both redundant patterns, that is, those patterns easy to detect and well represented by other patterns in the training set, and label noisy patterns, that is, those labelled as belonging to one class while their features clearly establish them as a member of another, while allowing to retain those patterns most interesting for training purposes. In section 4 we will illustrate numerically the results provided by the pattern selection algorithm over 7 example databases obtained fom the UCI repository. As we shall see, in all of them we arrive at much smaller training subsets that nevertheless allow the construction of effective PP classifiers. The paper ends with a brief summary section.

2 Parallel Perceptron Training

The parallel perceptron architecture is simply that of the well known committee machines. Let us briefly review it. Assume we are working with D dimensional patterns $X = (x_1, \ldots, x_D)^t$, where the D–th entry has a fixed 1 value to include bias effects. If the committee machine (CM) has H perceptrons, each with a weight vector W_i, for a given input X, the output of perceptron i is then $P_i(X) = s(W_i \cdot X) = s(act_i(X))$, where $s(\cdot)$ denotes the sign function and $act_i(X) = W_i \cdot X$ is the activation of perceptron i due to X. We then have

$$\begin{aligned}\sum_1^H P_i(X) &= \#\{i : W_i \cdot X > 0\} - \#\{i : W_i \cdot X < 0\} \\ &= N_+(X) - N_-(X) = \mathcal{N}(X),\end{aligned}$$

and the output $h(X)$ of the CM is $h(X) = s(\mathcal{N}(X))$ where we take H to be odd to avoid ties. We will assume that each input X has an associated ± 1 label l_X and take the output $h(X)$ as correct if $l_X h(X) > 0$. It is then clear that X has been correctly classified if either $N_+(X) > N_-(X)$ when $l_X = 1$ or $N_+(X) < N_-(X)$ when $l_X = -1$. If this is not the case, and we have, say, $l_X = 1$, then $N_-(X) = (H - \mathcal{N}(X))/2$. Classical CM training ([10], ch. 6) then tries to change the smallest number of perceptron outputs so that X could then be correctly classified, and it is easy to see that this number is $(1 + |\mathcal{N}(X)|)/2$; this last formula can also be applied to wrongly classified X such that $l_X = -1$. Then, whenever $l_X h(X) = -1$, classical CM training first selects those $(1 + |\mathcal{N}(X)|)/2$ perceptrons P_i such that $l_X P_i(X) = -1$ and for which $|act_i(X)|$ is smallest, and changes their weights by the well known Rosenblatt's rule:

$$W_i := W_i + \eta l_X X. \tag{1}$$

CMs and parallel perceptrons (PPs) differ in their training. PPs can be trained either on line or, as we shall do here, in batch mode and for them the update (1) is applied to all wrong perceptrons, i.e. those P_i verifying $l_X P_i(X) = -1$. Moreover, their training has a second ingredient, a margin–based output stabilization procedure. Notice that if $W_i \cdot X \simeq 0$, small changes on X may cause a wrong class assignment for a small perturbation of X. To avoid this instability, the update (1) is also applied when a pattern X is correctly classified but still $0 < l_X Act_i(X) < \gamma$.

The value of the margin γ is also adjusted dynamically from a starting value. More precisely, after a pattern X is processed correctly, we have

$$\gamma := \gamma + \eta \left(M_{min} - \min\{M_{max}, M(X)\} \right),$$

where $M(X)$ is the number of hyperplanes that process X correctly although with a too small margin. In other words, $M(X) = \#\{i : 0 < l_X Act_i(X) < \gamma\}$ for those X such that $l_X f(X) > 0$. Values proposed in [1] for M_{min} and M_{max} are $M_{min} = 0.25$ and $M_{max} = 1$. Observe then that γ increases if all correct perceptron activations are above the current margin (for then $M(X) = 0$), while it decreases if at least one perceptron activation is "below" the current margin (then $M(X) \geq 1$). Notice that for the margin to be meaningful, weights have to be normalized somehow; we will make its euclidean norm to be 1 after each batch pass. Notice PPs provide a common margin value γ for all H perceptrons; however, not all patterns have to behave with respect to γ in the same way overall H perceptrons.

In spite of their very simple structure, PPs do have a universal approximation property. Moreover, as shown in [1], PPs provide results in classification and regression problems quite close to those offered by procedures such as MLPs and C4.5 decision trees. Finally, their training is very fast, because of the very simple update (1) and because it is only applied to patterns incorrectly classified. Denoting their number as N_W and omitting for simplicity updates due to margins, the overall training complexity for a PP is $O(N_W DH)$; as training advances, we should have $N_W \ll N$ and, hence, very fast bacth iterations.

3 Training Pattern Selection

When dealing with imbalanced data sets, it is reasonable to expect patterns to fall within three cathegories, redundant, label noisy and borderline. Label noisy are simply those X for which their label assignment is likely to be wrong. It is thus desirable to exclude them from the training set. Redundant patterns are those easy to classify. Since they are likely to be overrepresented on the training set, many of them can be possibly ignored during training without hampering classifier construction. Finally, borderline patterns are those whose classification could be different after small perturbations and therefore, classifier construction should concentrate on them to provide stable and possibly correct classifications after training ends.

This training pattern cathegorization can be potentially quite useful, for once achieved, classifier construction can proceed by iteratively constructing a sequence of classifiers using training sets where redundant and noisy patterns are progressively removed. Notice that this training can be viewed as a kind of radical boosting–like procedure, where redundant and noisy patterns probabilities change to 0 after each iteration while the remaining patterns are taken as equiprobable. The difficulty obviously lies on how to characterize patterns beloging to each class. For this, activation margins are a natural choice. Recall that PPs adaptatively adjust this margin, making it to converge to a final value γ. If for a pattern X its i–th perceptron activation verifies $|act_i(X)| > \gamma$, it is likely to remain so after a small perturbation. Thus if for all i we have $l_X act_i(X) > \gamma$, X is likely to be also correctly classified later on. Those patterns are natural choices to be taken as redundant. Similarly, if for all i we have $l_X act_i(X) < -\gamma$, X is likely to remain wrongly classified, and we will take such patterns as label noisy. The remanining X will be the borderline patterns. We shall use the notations R_i, N_i and B_i for the redundant, noisy and borderline training sets at iteration i. With a slight abuse of the language, we shall call a pattern's normalized activation $l_X act_i(X)$ its "margin".

After iteration i we shall remove the R_i and N_i subsets. With respect to B_i the first option is to keep all of its patterns after each iteration. However, working with imbalanced data sets we should treat positive and negative training patterns in different ways, specially if classes are mixed. The alternative option we shall also use is to remove after each iteration the subset nB_i^- of "noisy" borderline negative patterns X with a wrong margin $l_X act_i(X) < 0$ in all perceptrons. Although we could also do the same for the noisy borderline positive set nB_i^+, this has the risk of removing a sizeable amount of the positive patterns, whose number could be much smaller than that of the negative ones. This second alternative option certainly favors the positive class, so it has to be balanced somehow. We shall do so with the termination criterium to be used. Recall that we want a good classification performance on the positive class, but maintaining also a good performance on the negative class. We thus need to balance the positive and negative accuracies, defined as $a^+ = TP/(TP + FN)$ and $a^- = TN/(TN + FP)$, where TP, TN denote the number of true positives and negatives, that is, positive and negative patterns correctly classified, and FP, FN denote the number of false positives and negatives, that is, negative and positive patterns incorrectly classified. For imbalanced problems, simple accuracy, that is the percentage of correctly classified patterns, may not be a relevant criterium, as it would be fulfilled by the simple (and very uninteresting) procedure of assigning all patterns to the (possibly much larger) negative classes. Other criteria are thus needed, and several options such as precision (the ratio $TP/(TP+FP)$) or recall (the ratio $TP/(TP+FN)$, i.e., our positive class accu-

Table 1. The table gives final g values when nB^- is kept (third column) and removed (fourth); the second option gives better results. For comparison purposes g values obtained after direct MLP are also given.

Problem set	% positives	final g with nB^-	final g without nB^-	MLP–BP
cancer	34.5	96.5	96.6	96.1
diabetes	34.9	68.7	71.2	71.7
ionosphere	35.9	77.6	82.3	80.8
vehicle	25.7	69.6	72.5	75.7
glass	13.6	91.2	91.6	91.8
vowel	9.1	92.9	93.3	97.1
thyroid	7.4	73.9	97.2	95.8

racy a^+), ROC curves, or other, have been proposed. Here we shall use a simple measure first used in [11], the geometric ratio $g = \sqrt{a^+a^-}$ between positive and negative accuracies, that measures the balance of the positive and negative class accuracies.

After the iterations end, the final PP is then used over the test set to determine the reported values of the overall test accuracy a_{ts} and the test set g_{ts} value, that will measure how well balanced are the generalization abilities of the just constructed classifier. The pseudocode of the general procedure including noisy borderline patterns is thus:

```
trSetReduction(trainingSet tr, testSet ts)
  gTr = 0;
  acc+_Tr = 0;
  trainPP(ts, g, acc+, W, gamma);       // first update of weigths, margin
  while g >= g_Tr and acc+ >= acc+_Tr:  // reduce Tr while g, acc+ improve
    gTr = g; acc+_Tr = acc+; wPP = W;   // W_PP: weights of best PP so far
    find(Tr, R, N, B, nB-, gamma);      // find redundant, l. noisy, borderline
    remove(tr, R, N);                   // remove redundant, label noisy
    remove(tr, nB-);                    // and negative noisy boderline
    trainPP(tr, g, acc+, W, gamma);
  calcAccG(ts, wPP, accTs, acc+_Ts, acc-_Ts, gTs);
```

In the next section we will illustrate numerically these procedures.

4 Numerical Results

We shall use 7 problem sets from the well known UCI database (listed in table 1) referring to the UCI database documentation [9] for more details on these problems. Some of them (glass, vowel, vehicle, thyroid) are multi–class problems; to reduce them to 2–class problems, we are taking as the minority classes the class 1 in the vehicle dataset, the class 0 in the vowel data set, and the class 7 in the glass domains (as done in [7]), and merged in a single class both sick thyroid classes. In general they can be considered relatively hard problems. Moreover, some of these problems provide well known examples of highly imbalanced positive and negative patterns, that difficult classifier construction, as discriminants

Table 2. Comparison of initial and final g values. The table also shows the training set reduction achieved.

Problem set	initial g	final g	initial Tr set	final Tr set	ave. # iters
cancer	96.8	96.6	629	174	1.99
diabetes	69.9	71.2	691	631	0.88
ionosphere	76.9	82.3	315	241	2.13
vehicle	67.0	72.5	762	284	2.89
glass	90.4	91.6	193	163	0.59
vowel	89.3	93.3	891	418	1.62
thyroid	68.1	97.2	6480	64	4.81

may tend to favor the (much) larger negative patterns over the less frequently positive ones. This is the case of the glass, vowel, thyroid and, to a lower extent, vehicle problems. In all of them we will take the minority class as the positive one.

PP training has been carried out as a batch procedure. In all examples we have used 3 perceptrons and parameters $\gamma = 0.05$ and $\eta = 10^{-2}$; for the thyroid dataset, we have taken $\eta = 10^{-3}$. As proposed in [1], the η rate does not change if the training error diminishes, but is decreased to 0.9η if it augments. Training epochs have been 250 in all cases; thus the training error evolution has not been taken into account to stop the training procedure. Anyway, it has an overall decreasing behavior. In all cases we have used 10–times 10–fold cross validation. That is, on each training stage, the overall data set has been randomly split in 10 subsets, 9 of which have been combined to obtain the initial training set, the size of which has been decreased on each training iteration as described above. To ensure an appropriate representation of positive pattern, stratified sampling has been used. The final PPs' behavior has been computed on the remaining, unchanged subset, that we keep for testing purposes.

Recall that we have discussed two handling options for training patterns in the set nB_i^-, either to keep or remove them. Table 1 gives the average of the final g values obtained over each test set. It also gives the proportion of positive patterns and the final g values given by a standard multilayer perceptron for comparison purposes. It can be seen that final g values arrived at removing patterns in nB^- are consistently better. Moreover, they favourably compare with MLP g values: although much simpler (and much faster to train), final PP g values are slightly better than MLP values in two cases, slightly worse in another two and essentially the same in the remaining three.

All other results will be given for training set selection when the nB_i^- sets are removed. They are contained in tables 2 and 3. The first table compares initial and final g values. In all cases but the cancer data set, final test g values are bigger than initial ones. The gain is small in some problems, that require few training set selection iterations, but much larger in other cases; the average number of iterations is nevertheless quite modest. For a quick comparison, we just mention that the g values for the vehicle and vowel problems are better than those in [7], where a different training set reduction method is used with

Table 3. Initial and final accuracy results for training set selection when patterns in nB^- are removed.

Problem set	initial acc	initial a^+	initial a^-	final acc	final a^+	final a^-
cancer	96.870	96.630	97.009	96.420	97.155	96.062
diabetes	75.039	57.926	84.369	75.158	61.373	82.665
ionosphere	82.143	64.051	92.367	86.171	71.803	94.255
vehicle	78.119	51.414	87.437	79.512	61.139	85.879
glass	95.048	84.500	96.795	95.842	86.000	97.459
vowel	97.273	80.667	98.933	97.838	88.111	98.811
thyroid	95.499	46.708	99.405	98.493	95.625	98.722

the 1–nearest neighbor (NN) and C4.5 algorithms; the glass g value reported here is slightly smaller than that reported there for the 1–NN method but better than that of C4.5 (notice that training sets used here may slightly differ from those used in [7]). On the other hand, except in the glass and diabetes problems training set reduction (shown in the same table) is quite marked, specially for the thyroid data set.

Table 3 compares initial and final accuracy values. In all cases final accuracy is bigger, except again for the cancer problem, where it remains essentially the same. As it should be expected, the algorithm enforced gain on the accuracy a^+ of the positive training class extends to the test sets, that show a noticeable increase, quite markedly in fact in all cases except the cancer dataset. On the other hand, the accuracy a^- of the negative class slightly increases in two cases, slightly decreases in another three and essentially stays the same in the remaining two cases.

As a summary of these results, we have illustrated that the proposed iterative training set selection procedure can achieve both noticeable improvements on the classification of a smaller positive class, while offering a good balance between positive and negative classification performances. Moreover, considerable reduction of training set sizes (and consequently a much faster training in the final iterations) are to be added to these advantages.

5 Conclusions and Further Work

In this paper we have proposed a new procedure for training set reduction based on the activation margins that arise naturally in parallel perceptron training. Its effectiveness has been verified on the seven 2–class problems studied here, several of them being representative of imbalanced class problems, where the discrimination of a small positive class may be damaged by the much larger number of negative samples. The proposed procedure balances in a natural way the number of positive and negative samples while ensuring a good generalization, not only in terms of a good overall test set accuracy, but also of its adequate balance among positive and negative classes.

This property, together with the very fast training of PP, may make them quite useful on large dimension imbalanced problems, an area of considerable interest as many interesting problems (text mining, microarray discrimination) belong to it. This and other questions, such as PP use in active training, and improvements in their performance, either by combining PPs through boosting or enlarging their parameter space, are under study.

References

1. P. Auer, H. Burgsteiner, W. Maass, *Reducing Communication for Distributed Learning in Neural Networks*, Proceedings of ICANN'2002, Lecture Notes in Computer Science 2415 (2002), 123–128.
2. L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, **Classification and Regression Trees**, Wadsworth, 1983.
3. N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer, *SMOTE: Synthetic Minority Oversampling Technique*, Journal of Artificial Intelligence Research 16 (2002), 321–357.
4. J. Dorronsoro, F. Ginel, C. Sánchez, C. Santa Cruz, *Neural Fraud Detection in Credit Card Operations*, IEEE Transactions on Neural Networks, 8 (1997), 827-834.
5. T. Fawcett, F. Provost, *Adaptive Fraud Detection*, Journal of Data Mining and Knowledge Discovery 1 (1997), 291–316.
6. Y. Freund *Boosting a weak learning algorithm by majority*, Information and Computation 121 (1995), 256–285.
7. M. Kubat, S. Matwin, *Addressing the Curse of Imbalanced Training Sets: One-Sided Selection*, Proceedings of the 14th International Conference on Machine Learning, ICML'97 (pp. 179-186), Nashville, TN, U.S.A.
8. M.A. Maloof, *Learning when data sets are imbalanced and when costs are unequal and unknown,* ICML-2003 Workshop on Learning from Imbalanced Data Sets II, 2003.
9. P. Murphy, D. Aha, *UCI Repository of Machine Learning Databases*, Tech. Report, University of Califonia, Irvine, 1994.
10. N. Nilsson, **The Mathematical Foundations of Learning Machines**, Morgan Kaufmann, 1990.
11. J.A. Swets, *Measuring the accuracy of diagnostic systems*, Science 240 (1998), 1285–1293.
12. G.M. Weiss, F. Provost, *The effect of class distribution on classifier learning*, Technical Report ML-TR 43, Department of Computer Science, Rutgers University, 2001.

Boosting Statistical Local Feature Based Classifiers for Face Recognition

Xiangsheng Huang and Yangsheng Wang

CASIA-SAIT HCI Joint Lab, Institute of Automation,
Chinese Academy of Sciences, Beijing, 10080, China
xiangshenghuang@hotmail.com

Abstract. In this work, we present a novel approach for face recognition which use boosted statistical local Gabor feature based classifiers. Firstly, two Gabor parts, real part and imaginary part, are extracted for each pixel of face images. The two parts are transformed into two kinds of Gabor features, magnitude feature and phase feature. 40 magnitude Gaborfaces and 40 phase Gaborfaces are generated for each face image by convoluting face images with five scales and eight orientations Gabor filters. Then these Gaborfaces are scanned with a sub-window from which the quantified Gabor features histograms are extracted representing efficiently the face image. The multi-class problem of face recognition is transformed into a two-class one of intra-and extra-class classification using intra-personal and extra-personal images, as in [5]. The intra/extra features are constructed based on these histograms of two different face images with Chi square statistic as dissimilarity measure. A strong classifier is learned using boosting examples, similar to the way in face detection framework [10]. Experiments on FERET database show good results comparable to the best one reported in literature [6].

1 Introduction

Face recognition has attracted much attention due to its potential values for applications as well as theoretical challenges. As a typical pattern recognition problem, face recognition has to deal with two main issues: (1) what features to use to represent a face, and (2) how to classify a new face image based on the chosen representation. Up until now, many representation approaches have been introduced, including Principal Component Analysis (PCA) [9], Linear Discriminant Analysis (LDA) [2], independent component analysis (ICA) [1], and Gabor wavelet features [11]. PCA computes a reduced set of orthogonal basis vector or eigenfaces of training face images. A new face image can be approximated by weighted sum of these eigenfaces. LDA seeks to find a linear transformation by maximising the between-class variance and minimising the within-class variance. ICA is a generalization of PCA, which is sensitive to the high-order relationships among the image pixels. Gabor wavelet captures the local structure corresponding to spatial frequency (scale), spatial localization, and orientation selectivity. Among various representations, multi-scale and multi-orientation Gabor features have attracted much attention and achieved great success in face recognition [7, 11].

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 51–58, 2005.

While regarding classification methods, nearest neighbor [9], convolutional neural networks, nearest feature line, Bayesian classification [5] and AdaBoost method have been widely used. The Bayesian Inra/Extraperson classifier (BIC) [5] uses the Bayesian decision theory to divide the difference vectors between pairs of face images into two classes: one representing intrapersonal differences (i.e. differences in a pair of image representing the same person) and extrapersonal differences.

Adaboost method, introduced by Freund and Schapire [3], which provides a simple yet effective stagewise learning approach for feature selection and nonlinear classification at the same time, has achieved great success in face detection [10] and other applications [7]. AdaBoost cascade framework, which is also widely used in many applications especially in face detection occasions [10], is a divide-and-conquer strategy, which can make training and testing process much easier and faster. Moreover, it is an efficient way to treat asymmetric problems and hardly converging training processes.

In this work, we present a novel approach for face recognition which use boosted statistical Gabor feature based classifiers. First two Gabor parts, real part and imaginary part, are extracted for each pixel of face images. The two parts are transformed into two kinds of Gabor features, magnitude feature and phase feature. 40 magnitude Gaborfaces and 40 phase Gaborfaces are generated for each face image by convoluting face images with five scales and eight orientations Gabor filters. Then these Gaborfaces are scanned with a sub-window from which the quantified Gabor features histograms are extracted representing efficiently the face image. The textures of the facial regions are locally encoded by the Gabor feature patterns while the shape of the face is recovered by the construction of the sub-window histogram. The idea behind using the local Gabor statistical features is that the face image can be seen as composition of micro-patterns which are invariant with respect to monotonic grey scale transformations. Combining these micro-patterns, a robust global description of the face image is obtained.

The multi-class problem of face recognition is transformed into a two-class one of intra-and extra-class classification using intra-personal and extra-personal images, as in [5]. The intra/extra features are constructed based on these histograms of two different face images with Chi square statistic as dissimilarity measure. A strong classifier is learned using boosting examples, similar to the way in face detection framework [10]. Experiments on FERET database show good results comparable to the best one reported in literature [6].

The rest of this paper is organized as follows: In section 2, the two kinds of Gabor feature face representation approach is introduced. The intra/extra features are constructed in section 3. In section 4, the cascade boosting learning for weak classifiers selection and classifier construction are proposed. And the experiment results using the FERET database and FERET evaluation protocol and analysis are shown in section 5. In section 6, we present the conclusion and future work.

2 Gabor Features for Face Image Representation

The representation of faces using Gabor feature has been extensively and successfully used in face recognition [4, 11]. Significant improvements in the face recognition rate have been reported in literature. Gabor features exhibit desirable characteristics of spatial locality and orientation selectively, and are optimally localized in the space and frequency domains. The Gabor kernels can be defined as follows:

$$\Psi_{\mu,\nu} = \frac{k_{\mu,\nu}^2}{\sigma^2} exp(-\frac{k_{\mu,\nu}^2 z^2}{2\sigma^2})[exp(ik_{\mu,\nu}z) - exp(-\frac{\sigma^2}{2})] \tag{1}$$

where μ and ν define the orientation and scale of the Gabor kernels, $z = (x, y)$ and the wave vector $k_{\mu,\nu}$ is defined as follows:

$$k_{\mu,\nu} = k_\nu e^{i\phi_\mu} \tag{2}$$

where $k_\nu = k_{max}/f^\nu, k_{max} = \pi/2, f = \sqrt{2}$, $\phi_\mu = 2\pi\mu/8$. The Gabor kernels in equ.(1) are all self-similar since they can be generated from one filter, the mother wavelet, by scaling and rotation via the wave vector $k_{\mu,\nu}$. We use Gabor kernels at five scales $\nu \in \{0, 1, 2, 3, 4\}$ and eight orientations $\mu \in \{0, 1, 2, 3, 4, 5, 6, 7\}$, with the parameter $\sigma = 2\pi$. The numbers of scales and directions selected made the feature extracted suitable to represent the characteristics of spatial locality and orientation selectivity as shown in Fig.1.

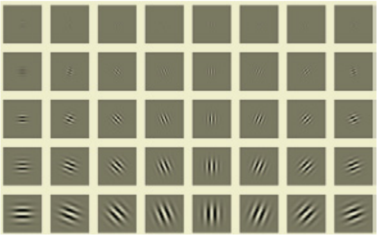


Fig. 1. 40 Gabor kernels used in this paper.

The Gaborfaces are computed by convoluting face images lixel we have two Gabor parts, real part and imaginary part:

Gabor real part:

$$Re(\Psi_{\mu,\nu}) = \frac{k_{\mu,\nu}^2}{\sigma^2} exp(-\frac{k_{\mu,\nu}^2 z^2}{2\sigma^2})[\cos(k_{\mu,\nu}z) - exp(-\frac{\sigma^2}{2})] \tag{3}$$

Gabor imaginary part:

$$Im(\Psi_{\mu,\nu}) = \frac{k_{\mu,\nu}^2}{\sigma^2} exp(-\frac{k_{\mu,\nu}^2 z^2}{2\sigma^2}) \sin(k_{\mu,\nu}z) \tag{4}$$

3 Intra/Extra Features Construction

3.1 Features Construction

Two Gabor parts, real part and imaginary part, are extracted for each pixel of face images. The two parts are transformed into two kinds of Gabor features, magnitude feature and phase feature:

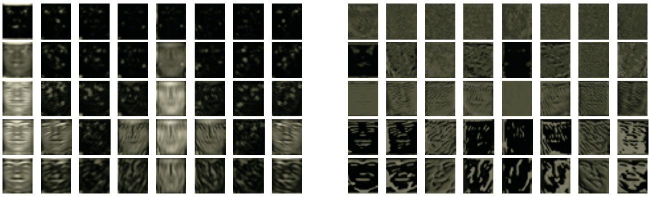
The magnitude features are generated by:

$$\sqrt{Re(\Psi_{\mu,\nu})^2 + Im(\Psi_{\mu,\nu})^2} \tag{5}$$

The phase features are calculated by:

$$\arctan(Re(Im(\Psi_{\mu,\nu})/Re(\Psi_{\mu,\nu})) \tag{6}$$

All feature values are quantified into 64 bins. 40 magnitude Gaborfaces and 40 phase Gaborfaces are generated for each face image by convoluting face image with five scales and eight orientations Gabor filters, as shown in Fig.2.



(a) 40 magnitude Gaborfaces (b) 40 phase Gaborfaces

Fig. 2. 80 Gaborfaces of one face image: 40 magnitude Gaborfaces and 40 phase Gaborfaces.

Then these Gaborfaces (in this work, face image size is 92×112) are scanned with a sub window which size is 20×20 by shifting two pixels at each step, from which the quantified Gabor features histograms are extracted. With the sub window moving, spatially enhanced feature histograms represent the face image efficiently. There are 1656 sub-windows for each Gaborface, and $132480(1656 \times 80)$ sub-windows for each face image. Each sub-window has one histogram to describe the local statistical property of Gabor feature. The multi-class problem of face recognition is transformed into a two-class one of intra-and extra-class classification. The basic idea is that intra-personal images have similar local Gabor feature statistical property, in contrast, extra-personal images have dissimilar local statistical property. In this work, Chi square statistic is applied to two corresponding histograms of two face images to measure the similarity of the two face images. That is to say, there are intra/extra 132480 features in the statistical Gabor feature space for each pair of face images.

3.2 Chi Square Statistic Distance

A histogram of the Gaborface $f_l(x,y)$ can be defined as:

$$H_j = \sum_{x,y} I\{f_l(x,y) = j\}, \qquad j = 0, \ldots, n-1 \tag{7}$$

in which n is the number of different labels produced by the Gabor filters (In this work, Gabor coefficients are quantified into 64 bins, so n is 64) and

$$I\{A\} = \begin{cases} 1, & A \; is \; true \\ 0, & A \; is \; false \end{cases} \tag{8}$$

This histogram contains information about the distribution of the local micro-patterns, such as edges, spots and flat areas, over the whole Gaborface. For efficient face representation, one should retain also spatial information. For this purpose, the Gaborface is scanned with a sub-window and the spatially enhanced histogram of each window, W_i, is defined as:

$$H_{i,j} = \sum_{x,y} I\{(x,y) \in W_i\} I\{f_l(x,y) = j\} \tag{9}$$

In this histogram, we effectively have a description of the face on two different levels of locality: the labels for the histogram contain information about the patterns in pixel-level, and the labels are summed over a small region to produce information on regional level.

Several possible dissimilarity measures have been proposed for histograms. In this work, Chi square statistic is adopted.

-Chi square statistic (χ^2):

$$\chi^2(S,M) = \sum_i \frac{(S_i - M_i)^2}{(S_i + M_i)} \tag{10}$$

When the image has been divided into regions, it can be expected that some of the regions contain more useful information than others in terms of distinguishing between people. For examples, eyes seem to be an important cue in human face recognition. To take advantage of this, AdaBoost is applied to select intra/extra feature and set them with different weight based on the importance of the information it contains.

4 Feature Selection and Classifier Learning

The set of 132480 intra/extra features is an over-complete set for the intrinsically low dimensional face appearance pattern and contains much redundant information. We propose to use Adaboost to select most significant intra/extra features from a large feature set.

Therefore, AdaBoost is adopted to solving the following three fundamental problems in one boosting procedure: (1) learning effective features from a large

```
0. (Input)
    (1) Training examples $\{(x_1, y_1), \ldots, (x_N, y_N)\}$,
      where $N = a + b$; of which $a$ examples have $y_i = +1$
      and $b$ examples have $y_i = -1$;
1. (Initialization)
    $w_{0,i} = \frac{1}{2a}$ for those examples with $y_i = +1$ or
    $w_{0,i} = \frac{1}{2b}$ for those examples with $y_i = -1$.
2. (Forward Inclusion)
  For $t = 1, \ldots, T$
    (1) Train one hypothesis $h_j$ for each feature $j$ with $w_t$, and
      error $e_j = Pr_i^{w_t}[h_j(x_i) \neq y_i]$
    (2) Choose $h_t(x) = h_k(x)$, such that $\forall j \neq k$, $e_k < e_j$.
      Let $e_t = e_k$.
    (3) Update $w_{t+1,i} \leftarrow w_{t,i}\beta_t^{I_i}$, where $I_i = 1$ or 0 for example
      $x_i$ classified correctly or incorrectly respectively and
      $\beta_t = e_t/(1 - e_t)$, normalize to $\sum_i w_{t+1,i} = 1$;
3. (Output)
    $H(x) = \begin{cases} 1, if & \sum_{t=1}^{T} \alpha_t h_t(x) > \sum_{t=1}^{T} \alpha_t \\ 0, & otherwose \end{cases}$
where $\alpha_t = \log \frac{1}{\beta_t}$
```

Fig. 3. AdaBoost Algorithm.

feature set, (2) constructing weak classifiers each of which is based on one of the selected features, and (3) boosting the weak classifiers into a stronger classifier.

The AdaBoost algorithm based on the descriptions from [4, 8] is shown in Fig. 3. The AdaBoost learning procedure is aimed to derive α_t and $h_t(x)$. Every training example is associated with a weight. During the learning process, the weights are updated dynamically in such a way that more emphasis is placed on hard examples which are erroneously classified previously.

5 Experiments

We tested the proposed method on the FERET $fafb$ face database, and the training set is also from the training set of the FERET database, which includes 1002 images of 429 subjects. All images are cropped to 112 pixels high by 92 pixels wide and rectified according to the eye positions provided with the FERET data. Histogram normalization is used to preprocess all cropped images. The cropped and preprocessed images are illustrated in Fig.4. 795 intra-personal image pairs and $500,706$ extra-personal image pairs are generated using the training set.

To test the efficiency of our proposed method, several comparative experiments were tested on the probe set fb with the gallery fa of the FERET database. There are 1196 images in fa, 1195 images in fb, and all exactly the

Fig. 4. Some examples of preprocessed face images.

subjects have exactly one image in both fa and fb. The rank curves of the final recognition results are plotted in Fig.5. It should be noted that the CSU implementations of the algorithms whose results we introcued here do not achieve the same figures as in original FERET test due to some modifications in the experimental setup. Our approach has achieved the upper bound recognition performance shown in Fig.5.

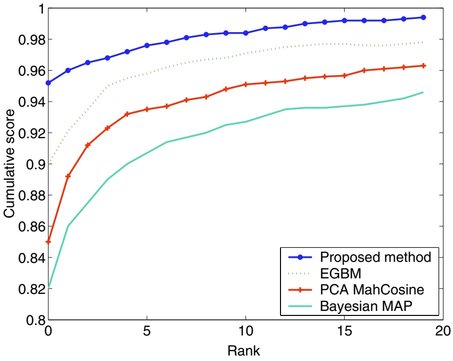


Fig. 5. Rank curves for the fb probe sets.

6 Conclusion

In this work, we present a novel approach for face recognition which use boosted statistical local Gabor feature based classifiers. The textures of the facial regions are locally encoded by the Gabor feature patterns while the shape of the face is recovered by the construction of the sub-window histogram. The idea behind using the local Gabor statistical features is that the face image can be seen as composition of micro-patterns which are invariant with respect to monotonic grey scale transformations. Combining these micro-patterns, a robust global description of the face image is obtained. The multi-class problem of face recognition is transformed into a two-class one of intra-and extra-class classification using intra-personal and extra-personal images, as in [5]. The intra/extra features are constructed based on these histograms of two different face images with Chi square statistic as dissimilarity measure. A strong classifier is learned using boosting examples. Experimental results on FERET(fa fb)database has proven the effectiveness of our new approach. While the problem of how to select the kernel size of Gabor filter and scanning sub-window size is still open. And this will be the focus in our future research.

References

1. Marian Stewart Bartlett, H. Martin Lades, and T. J. Sejnowski. Independent component representations for face recognition. *Proceedings of the SPIE, Conference on Human Vision and Electronic Imaging III*, 3299:528–539, 1998.
2. K. Etemad and R. Chellapa. “Face recognition using discriminant eigenvectors”. In *Proceedings of the International Conference on Acoustic, Speech and Signal Processing*, 1996.
3. Y. Freund and R.E. Schapire. “A decision-theoretic generalization of on-line learning and an application to boosting”. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997.
4. J. Friedman, T. Hastie, and R. Tibshirani. “Additive logistic regression: a statistical view of boosting”. *The Annals of Statistics*, 28(2):337–374, April 2000.
5. B. Moghaddam, C. Nastar, and A. Pentland. “A Bayesain similarity measure for direct image matching”. *Media Lab Tech Report* No. 393, MIT, August 1996.
6. P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.
7. L.Zhang S.Z.Li Z.Y. Qu and X.S.Huang. “Boosting local feature based classifiers for face recognition”. In *The First IEEE Workshop on Face Processing in Video*, Washington D.C, June 2004.
8. R. E. Schapire and Y. Singer. “Improved boosting algorithms using confidence-rated predictions”. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 80–91, 1998.
9. Matthew A. Turk and Alex P. Pentland. “Eigenfaces for recognition”. *Journal of Cognitive Neuroscience*, 3(1):71–86, March 1991.
10. P. Viola and M. Jones. “Robust real time object detection”. In *IEEE ICCV Workshop on Statistical and Computational Theories of Vision*, Vancouver, Canada, July 13 2001.
11. L. Wiskott, J.M. Fellous, N. Kruger, and C. Vonder malsburg. ”face recognition by elastic bunch graph matching”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.

Dynamic and Static Weighting in Classifier Fusion

Rosa M. Valdovinos[1], J. Salvador Sánchez[2], and Ricardo Barandela[1]

[1] Instituto Tecnológico de Toluca, Av. Tecnológico s/n, 52140 Metepec, México
{li_rmvr,rbarandela}@hotmail.com
[2] Dept. Llenguatges i Sistemes Informàtics, Universitat Jaume I, 12071 Castelló, Spain
sanchez@uji.es

Abstract. When a Multiple Classifier System is employed, one of the most popular methods to accomplish the classifier fusion is the simple majority voting. However, when the performance of the ensemble members is not uniform, the efficiency of this type of voting is affected negatively. In this paper, a comparison between simple and weighted voting (both dynamic and static) is presented. New weighting methods, mainly in the direction of the dynamic approach, are also introduced. Experimental results with several real-problem data sets demonstrate the advantages of the weighting strategies over the simple voting scheme. When comparing the dynamic and the static approaches, results show that the dynamic weighting is superior to the static strategy in terms of classification accuracy.

1 Introduction

A multiple classifier system (MCS) is a set of individual classifiers whose decisions are combined when classifying new patterns. There are many different reasons for combining multiple classifiers to solve a given learning problem [6], [12]. First, MCSs try to exploit the local different behavior of the individual classifiers to improve the accuracy of the overall system. Second, in some cases MCS might not be better than the single best classifier but can diminish or eliminate the risk of picking an inadequate single classifier. Another reason for using MCS arises from the limited representational capability of learning algorithms. It is possible that the classifier space considered for the problem does not contain the optimal classifier.

Let D = { D_1, ..., D_h } be a set of classifiers. Each classifier assigns an input feature vector $\mathbf{x} \in \Re^n$ to one of the c problem classes. The output of a MCS is an h-dimensional vector containing the decisions of each of the h individual classifiers:

$$[D_1(\mathbf{x}),..., D_h(\mathbf{x})]^T \quad (1)$$

It is accepted that there are two main strategies in combining classifiers: selection and fusion. In classifier selection, each individual classifier is supposed to be an expert in a part of the feature space and therefore, we select only one classifier to label the input vector **x**. In classifier fusion, each component is supposed to have knowledge of the whole feature space and correspondingly, all individual classifiers decide the label of the input vector.

Focusing on the fusion strategy, the combination can be made in many different ways. The simplest one employs the majority rule in a plain voting system [4]. More

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 59–66, 2005.

elaborated schemes use weighted voting rules, in which each individual component is associated with a different weight [5]. The final decision can be made by majority, average [6], minority, medium [7], product of votes, or using some other more complex methods [8], [9], [10], [19].

In the present work, some methods for weighting the individual components in a MCS are proposed, and their effectiveness is empirically tested over real data sets. Three of these methods correspond to the so-called dynamic weighting, by using the distances to a pattern. The last method, which belongs to the static weighting strategy, estimates the leaving-one-out error produced by each classifier in order to set the weights of each component [21].

From now on, the rest of the paper is organized as follows. Sect. 2 provides a brief review of the main issues related to classifier fusion and makes a very simple categorization of weighting methods, distinguishing between dynamic and static weighting of classifiers. Moreover, seveal weighting procedures are also introduced in Sect. 2. The experimental results are discussed in Sect. 3. Finally, some conclusions and possible further extensions are given in Sect. 4.

2 Classifier Fusion

As pointed out in Sect. 1, classifier fusion assumes that all individual classifiers are competitive, instead of complementary. For this reason, each component takes part in the decision of classifying an input test pattern.

In the simple voting (by majority), the final decision is taken according to the number of votes given by the individual classifiers to each one of the classes, thus assigning the test pattern to the class that has obtained a majority of votes. When working with data sets that contain more than two classes, in the final decision ties among some classes are very frequently obtained. To solve this problem, several criteria can be considered. For instance, to randomly take the decision, or to implement an additional classifier whose ultimate goal is to bias the decision toward a certain class [15].

An important issue that has strongly called the attention of many researchers is the error rate associated to the simple voting method and to the individual components of a MCS. Hansen and Salomon [17] show that if each one of the classifiers being combined has an error rate less than 50%, it may be expected that the accuracy of the ensemble improve when more components are added to the system. However, this assumption not always is fulfilled. In this context, Matan [18] asserts that in some cases, the simple voting might perform even worse than any of the members of the MCS. Thus some weighting method can be employed in order to partially overcome these difficulties.

A weighted voting method has the potential to make the MCS more robust to the choice of the number of individual classifiers. Two general approaches to weighting can be remarked: dynamic weighting and static weighting of classifiers. In the dynamic strategy, the weights assigned to the individual classifiers can change for each test pattern. On the contrary, in the static weighting, the weights are computed for each classifier in the training phase, and they are maintained constant during the classification of the test patterns.

In the following sections, several weighting functions, both from the dynamic and the static categories, are explored. It has to be noted that in the present work, all the individual classifiers correspond to the 1-NN (Nearest Neighbor) rule [16]. This is a well-known supervised non-parametric classifier that combines conceptual and implementational simplicity with an asymptotic error rate conveniently bounded in terms of the optimal Bayes error. In its classical manifestation, given a set of *m* previously labeled instances (or training set, TS), this classifier assigns any input test pattern to the class indicated by the label of the closest example in the TS. The extension of this rule corresponds to the *k*-NN classifier, which consists of assigning an input pattern to the class most frequently represented among the *k* closest training instances.

2.1 Dudani's Dynamic Weighting

A weighted *k*-NN rule for classifying new patterns was first proposed by Dudani [3]. The votes of the *k* nearest neighbors are weighted by a function of their distance to the test pattern. In his original proposal, a neighbor with smaller distance is weighted more heavily than one with a greater distance: the nearest neighbor gets a weight of 1, the furthest neighbor a weight of 0, and the other weights are scaled linearly to the interval in between (Eq. 2):

$$w_j = \begin{cases} \dfrac{d_k - d_j}{d_k - d_1} & \text{if } d_k \neq d_1 \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

where d_j denotes the distance of the *j*'th nearest neighbor to the test pattern, d_1 is the distance of the nearest neighbor, and d_k indicates the distance of the furthest (*k*'th) neighbor.

Now, this function will be here applied to make the dynamic weighting of the individual components in an ensemble. Correspondingly, the value of *k* (that is, the number of nearest neighbors in Dudani's rule) will be replaced by the number of classifiers *h* that constitute the MCS. The procedure to assign the weights can be described as follows:

```
1. Let $d_j$ ($j$ = 1, …, $h$) be the distance of an input test vector x to its nearest neighbor in the $j$'th individual classifier.
2. Sort the $h$ distances in increasing order: $d_1$, …, $d_h$.
3. Weight classifier $D_j$ by means of function in Eq. 2.
```

2.2 Dynamic Weighting by Index

Another weighting function is here considered. Like in Dudani's method, the *h* distances of the test pattern **x** to its nearest neighbor in each individual classifier have also to be sorted. In this case, each classifier D_j is weighted according to the following function:

$$w_j = h - j + 1 \tag{3}$$

where *j* represents the index of an individual classifier after sorting the corresponding *h* distances.

Consider a MCS consisting of three individual classifiers D = {D_1, D_2, D_3}. The distance of the nearest neighbor to a given test pattern **x** by means of each classifier is d_1, d_2, and d_3, respectively. Now suppose that $d_2 < d_1 < d_3$. Thus after sorting the three distances, the index of classifier D_1 is 2, the index of D_2 is 1, and the index of D_3 is 3. Consequently, by applying the weighting function in Eq. 3, the resulting weights are $w_1 = 3 - 2 + 1 = 2$, $w_2 = 3 - 1 + 1 = 3$, and $w_3 = 3 - 3 +1 = 1$.

2.3 Dynamic Weighting by Averaged Distances

We here propose a novel weighting function, which is based on the computation of averaged distances. In summary, the aim of this new dynamic weighting procedure is to reward (by assigning the highest weight) the individual classifier with the nearest neighbor to the input test pattern. The rationale behind this is that such a classifier probably corresponds to that with the highest accuracy in the classification of the given test pattern. Thus each classifier D_j will be weighted by means of the function shown in Eq. 4:

$$w_j = \frac{\sum_{i=1}^{h} d_i}{d_j} \tag{4}$$

Note that, by using this weighting function, we effectively accomplish the goal previously stated, that is, the individual classifier with the smallest distance will get the highest weight, while the one with the greatest distance will obtain the lowest weight.

2.4 Static Weighting by Leaving-One-Out Error Estimate

While the previous methods weight the individual components of a MCS in a dynamic manner, the last proposal corresponds to the static category. In this sense, weighting will be here performed in the training phase by means of the leaving-one-out error estimate method. To this end, for each individual classifier D*j*, the following function e_j is defined:

$$e_j = \frac{1}{m} \sum_{x \in S} e(y, x) \tag{5}$$

where m denotes the number of patterns in a training sample S, x represents a training instance, y is the nearest neighbor of x in $S - \{x\}$, and $e(y, x)$ is defined as follows:

$$e(y,x) = \begin{cases} 0 & \text{if } \; L(y) = L(x) \\ 1 & \text{otherwise} \end{cases} \tag{6}$$

where $L(x)$ is the class label of a pattern x, and $L(y)$ indicates the class label of a pattern y.

By using the error function just introduced, each individual classifier D_j will be weighted according to the function in Eq. 7:

$$w_j = 1 - \frac{e_j / m}{\sum_{i=1}^{h} e_i} \tag{7}$$

Note that this weight is directly related to the amount of errors produced by each individual classifier. Thus the classifier with the smallest error will be assigned the highest weight, while the one with the greatest error will obtain the lowest weight.

3 Experimental Results

The results here reported correspond to the experiments over six real data sets taken from the UCI Machine Learning Database Repository [11]. For each data set, the 5-fold cross-validation error estimate method was employed: 80% of the available patterns were for training purposes and 20% for the test set.

The integration of the MCS was performed by manipulating the patterns [12] for each of the classes, thus obtaining three different individual classifiers with four variants:

- Sequential selection [1], [2] (Sel1)
- Random selection with no replacement [1], [2] (Sel2)
- Selection with Bagging [13] (Sel3)
- Selection with Boosting [14] (Sel4)

The experimental results given in Table 1 correspond to the averages of the general accuracy in the fusion, by technique of pattern selection and method of weighting. The 1-NN classification accuracy for each entire original TS (i.e., with no combination) has also been included as the baseline classifier. Analogously, the results for the MCS with simple voting (no weighting) are reported for comparison purposes.

From results in Table 1, some preliminary conclusions can be drawn. First, for all data sets there exists at least one classifier fusion technique whose classification accuracy is higher than that obtained when using the whole TS (i.e., with no combination). Second, comparing the four selection methods, in general Sel1 and Sel4 clearly outperform the other two selection approaches (namely, random with no replacement and bagging), independent of the voting scheme adopted. On the other hand, focusing on sequential selection (Sel1) and boosting (Sel4), the accuracy of Sel1 results superior to that of Sel4 in most cases (22 out of 30).

If we now compare the simple and the weighted voting schemes, we can observe that in all data sets, we can find a weighting technique with better results than those of the simple majority voting. The Dudani's weighting outperforms all the other methods in Liver database. The weighting by index is the best in Cancer and Glass domains. The weighting by averaged distances achieves the highest accuracy in Heart, Pima and Vehicle databases.

Finally, with respect to differences in accuracy between dynamic and static weighting, it has to be especially remarked the fact that results of the static strategy are always inferior to those of the dynamic approach. As can be seen, although differences are not significant, the static weighting does not seem to present any advantage with respect to the dynamic weightings.

Table 1. Averaged accuracy of different classifier fusion methods. Values in italics indicate the best selection method for each voting scheme and each data set. Boldface is used to emphasize the highest accuracy for each problem

	Cancer	Heart	Liver	Pima	Glass	Vehicle
Original TS	95.62	58.15	65.22	65.88	70.00	64.24
Simple voting						
Sel1	*96.93*	*65.19*	*63.77*	68.89	*68.00*	*64.48*
Sel2	66.42	50.37	57.10	59.35	56.50	62.10
Sel3	72.12	45.19	50.14	60.00	60.50	60.55
Sel4	94.16	57.78	62.03	*70.07*	62.50	60.43
Dudani's weighting						
Sel1	95.62	58.15	***65.51***	*68.37*	*70.00*	*64.24*
Sel2	68.47	52.96	56.23	59.08	67.00	61.02
Sel3	74.16	47.41	52.17	60.26	65.00	60.91
Sel4	*95.89*	*58.52*	60.87	67.58	66.50	64.24
Weighting by index						
Sel1	95.91	*61.11*	*62.61*	68.24	***71.00***	*64.48*
Sel2	65.84	54.07	53.04	62.09	62.00	62.34
Sel3	72.41	47.78	49.28	60.92	61.50	60.79
Sel4	***99.27***	57.41	59.42	*70.07*	66.00	62.81
Weighting by averaged distances						
Sel1	*96.50*	***65.56***	*65.22*	68.37	*68.00*	***64.72***
Sel2	62.04	49.63	57.10	59.08	59.00	59.00
Sel3	70.80	45.93	50.14	60.26	62.50	63.41
Sel4	93.58	57.78	62.32	***70.85***	63.00	61.50
Static weighting						
Sel1	*96.93*	*65.19*	*63.77*	68.89	*68.50*	*63.65*
Sel2	66.42	50.37	57.10	59.35	56.00	62.93
Sel3	72.12	45.19	50.14	60.00	60.50	59.84
Sel4	94.16	59.63	62.03	*70.07*	63.00	61.03

4 Conclusions and Future Work

In a MCS, performance mainly depends on the accuracy of the individual classifiers and on the specific way of combining the individual decisions. Correspondingly, it results crucial to appropriately handle the combination of decisions in order to attain the most accurate system. In the present work, several weighting methods, both from the dynamic and static approaches, have been introduced and empirically compared with the simple majority voting scheme.

From the experiments carried out, our study shows that the weighting voting clearly outperforms the simple voting procedure, which erroneously assumes the uniform performance of the individual components of a MCS. Another issue to remark is that the dynamic weighting is superior to the static strategy, in terms of classification accuracy.

At this moment, it has to be admitted that it results difficult enough to propose one of the dynamic weightings as the best method. In fact, differences among them are more or less significant depending on each particular database. Nevertheless, one can see that the weighting by averaged distances achieves the highest accuracy in 3 out of 6 problems (50% of the cases), while the weighting by index in 2 out of 6 databases (33% of the cases).

Future work is primarily addressed to investigate other weighting functions applied to classifier fusion. For instance, the inverse distance function proposed by Shepard [20] could represent a good alternative to other weighted voting schemes with low classification accuracy. On the other hand, the results reported in this paper should be viewed as a first step towards a more complete understanding of the behavior of the weighted voting procedures and consequently, it is still necessary to perform a more extensive analysis of the dynamic and static weighting strategies over a larger number of synthetic and real problems.

Acknowledgements

This work has been partially supported by the Programme Alβan, the European Union Programme of High Level Scholarships for Latin America (Scholarship No. E04D033674MX), and by grants TIC2003-08496 from the Spanish CICYT, GV04A/705 from Generalitat Valenciana, and P1-1B2004-08 from Fundació Caixa Castelló-Bancaixa.

References

1. Barandela, R., Valdovinos, R.M., Sánchez, J.S.: New applications of ensembles of classifiers, Pattern Analysis and Applications 6 (2003) 245-256.
2. Valdovinos, R.M., Barandela, R.: Sistema de Múltiples Clasificadores. Una alternativa para la Escalabilidad de Algoritmos, In: Proc. of the 9th Intl. Conference of Research on Computer Sciences, Puebla, Mexico (2002).
3. Dudani, S.A.: The distance weighted k-nearest neighbor rule, IEEE Trans. on Systems, Man and Cybernetics 6 (1976) 325-327.
4. Kuncheva, L.I., Kountchev, R.K.: Generating classifier outputs of fixed accuracy and diversity, Pattern Recognition Letters 23 (2002) 593–600.
5. Woods, K., Kegelmeyer Jr., W.P, Bowyer, K.,: Combination of multiple classifiers using local accuracy estimates, IEEE Trans. on Pattern Analysis and Machine Intelligence 19 (1997) 405-410.
6. Kuncheva, L.I.: Using measures of similarity and inclusion for multiple classifier fusion by decision templates, Fuzzy Sets and Systems 122 (2001) 401-407.
7. Chen, D., Cheng, X.: An asymptotic analysis of some expert fusion methods, Pattern Recognition Letters 22 (2001) 901–904.
8. Kuncheva, L.I., Bezdek, J.C., Duin, R.P.W.: Decision templates for multiple classifier fusion, Pattern Recognition 34 (2001) 299-314.
9. Ho, T.-K.: Complexity of classification problems and comparative advantages of combined classifiers, In: Proc. of the 1st Intl. Workshop on Multiple Classifier Systems, Springer (2000) 97-106.
10. Bahler, D., Navarro, L.: Methods for combining heterogeneous sets of classifiers, In: Proc. of the 17th Natl. Conference on Artificial Intelligence (AAAI-2000), Workshop on New Research Problems for Machine Learning (2000).
11. Merz, C.J., Murphy, P.M.: UCI Repository of Machine Learning Databases, Dept. of Information and Computer Science, Univ. of California, Irvine, CA (1998).

12. Dietterich, G.T.: Machine learning research: four current directions, AI Magazine 18 (1997) 97–136.
13. Breiman, L.: Bagging predictors, Machine Learning 24 (1996) 123-140.
14. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm, In: Proc. of the 13th Intl. Conference on Machine Learning, Morgan Kaufmann (1996) 148-156.
15. Kubat, M., Cooperson Jr., M.: Voting nearest neighbor subclassifiers, In: Proc. of the 17th Intl. Conference on Machine Learning, Morgan Kaufmann, Stanford, CA (2000) 503-510.
16. Dasaraty, B.V..: Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques, IEEE Computer Society press, Los Alamitos, CA (1991).
17. Hansen, L.K., Salomon, P. : Neural network ensembles, IEEE Trans. on Pattern Analysis and Machine Intelligence 12 (1990) 993-1001.
18. Matan, O.: On voting ensembles of classifiers, In: Proc. of the 13th Natl. Conference on Artificial Intelligence (AAAI-96), Workshop on Integrating Multiple Learned Models (1996) 84–88.
19. Ho, T.-K., Hull, J.J., Srihari, S.N.: Combination of Decisions by Multiple Classifiers, Structured Document Image Analysis, In: Springer-Verlag, Heidelberg (1992) 188–202.
20. Shepard, R.N.: Toward a universal law of generalization for psychological science, Science 237 (1987) 1317-1323.
21. Verikasa, A., Lipnickasb A., Malmqvista, K., Bacauskieneb, M., Gelzinisb, A.: Soft combination of neural classifiers: a comparative study, Pattern Recognition Letters 20 (1999) 429-444.

A Novel One-Parameter Regularized Kernel Fisher Discriminant Method for Face Recognition

Wensheng Chen[1], Pongchi Yuen[2], Jian Huang[2], and Daoqing Dai[3]

[1] Department of Mathematics, Shenzhen University, P.R. China, 518060
chenws@szu.edu.cn
[2] Department of Computer Science, Hong Kong Baptist University
{jhuang,pcyuen}@comp.hkbu.edu.hk
[3] Department of Mathematics, Sun Yat-sen University, P.R. China
stsddq@zsu.edu.cn

Abstract. Kernel-based regularization discriminant analysis (KRDA) is one of the promising approaches for solving small sample size problem in face recognition. This paper addresses the problem in regularization parameter reduction in KRDA. From computational complexity point of view, our goal is to develop a KRDA algorithm with minimum number of parameters, in which regularization process can be fully controlled. Along this line, we have developed a Kernel 1-parameter RDA (K1PRDA) algorithm (W. S. Chen, P C Yuen, J Huang and D. Q. Dai, "Kernel machine-based one-parameter regularized Fisher discriminant method for face recognition," *IEEE Transactions on SMC-B*, to appear, 2005.). K1PRDA was developed based on a three-parameter regularization formula. In this paper, we propose another approach to formulate the one-parameter KRDA (1PRKFD) based on a two-parameter formula. Yale B database, with pose and illumination variations, is used to compare the performance of 1PRKFD algorithm, K1PRDA algorithm and other LDA-based algorithms. Experimental results show that both 1PRKFD and K1PRDA algorithms outperform the other LDA-based face recognition algorithms. The performance between 1PRKFD and K1PRDA algorithms are comparable. This concludes that our methodology in deriving the one-parameter KRDA is stable.

1 Introduction

Among various appearance-based techniques, linear discriminant analysis-based (LDA) method is one of the promising approaches in face recognition. The first well-known LDA-based face recognition algorithm is so-called Fisherface [1] developed in 1997. However, LDA has two major drawbacks for its application to Pattern Recognition (PR). First, the distributions of face image variations under different pose and illumination is complex and nonlinear. Therefore, like other appearance-based methods, the performance of LDA-based method will degrade under pose and illumination variations. To overcome this drawback,

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 67–74, 2005.

kernel method is employed. The basic idea is to apply a nonlinear mapping $\Phi : x \in R^d \rightarrow \Phi(x) \in F$ to the input data vector x in input space R^d and then to perform the LDA on the mapped feature space F. This method is socalled Kernel Fisher Discriminant (KFD) [2]. Secondly, many LDA-based algorithms usually suffer from small sample size (S3) problem. Some algorithms, such as Fisherface [1], Direct LDA [3] and RDA [4, 6] etc, are developed to solve S3 problem. However, Fisherface and Direct LDA are implemented in the sub-feature-space, not in the full feature space. So it may lost some useful discriminant information in projection to a subspace. Dai et al. [4] proposed three parameters regularized method to solve S3 problem. Although this method is executed in the full sample space, it's very difficulty to determine three optimal parameters. We have proposed reducing to three parameters to one parameter and developed a Kernal 1-parameter RDA (K1PRDA) method [6]. The results are encouraging. In this paper, we propose and develop another 1-parameter RKFD (1PKFD)algorithm. However, the starting point is not a 3-parameter formulation, but a 2-parameter formulation [5]. The optimal parameters (θ, t) are determined simultaneously by using techniques proposed in [6].

2 Proposed Method

This paper proposes and develops another one parameter regularization KFD method for face recognition. Details are discussed as follows.

2.1 Some Definitions

Assume the dimensionality of original sample feature space be d and the number of sample classes be C, the total original sample $X = \{X_1, X_2, \cdots, X_C\}$, the jth class X_j contains N_j samples, namely $X_j = \{x_1^j, x_2^j, \cdots x_{N_j}^j\}$, $j = 1, 2, \cdots, C$. Let N be the total number of original training samples, so $N = \sum_{j=1}^{C} N_j$. Let nonlinear mapping $\Phi : x \in R^d \rightarrow \Phi(x) \in F$, where F is the mapped feature space, denote df=dim F. Let $m_j = \frac{1}{N_j}\sum_{x\in X_j}\Phi(x)$ be the mean of the mapped sample class $\Phi(X_j)$ and $m = \frac{1}{N}\sum_{j=1}^{C}\sum_{x\in X_j}\Phi(x)$ be the global mean of the total mapped sample $\Phi(X)$. The matrices S_w^Φ, S_b^Φ are defined respectively as:

$$S_w^\Phi = \frac{1}{N}\sum_{j=1}^{C}\sum_{x\in X_j}\left(\Phi(x) - m_j\right)\left(\Phi(x) - m_j\right)^T, \; S_b^\Phi$$

$$= \frac{1}{N}\sum_{j=1}^{C} N_j(m_j - m)(m_j - m)^T$$

The Fisher index $J_\Phi(w)$ in F is defined as

$$J_\Phi(w) = \frac{w^T S_b^\Phi w}{w^T S_w^\Phi w}, \quad w \in F. \tag{1}$$

2.2 Kernel Fisher Discriminant Analysis (KFDA)

According to Mercer kernel function theory [7], any solution $w \in F$ must belong to the span of all training patterns in F. Hence there exists a group of constants $\{\tilde{w}_k^l\}_{1 \le l \le C, 1 \le k \le N_l}$ such that $w = \sum_{l=1}^{C} \sum_{k=1}^{N_l} \tilde{w}_k^l \Phi(x_k^l)$. If substituting w into (1), it yields that the Fisher criterion function in the mapped feature space F can be written as followings:

$$J_\Phi(\tilde{w}) = \frac{\tilde{w}^T P_\Phi \tilde{w}}{\tilde{w}^T Q_\Phi \tilde{w}} \tag{2}$$

where $\tilde{w} = (\tilde{w}_k^l)_{1 \le l \le C, 1 \le k \le N_l} \in R^N$.

LDA is to solve the problem $\tilde{w}^* = \arg\max_{\tilde{w} \in F} J_\Phi(\tilde{w})$, which is equivalent to solving eigenvalue problem $(Q_\Phi^{-1} P_\Phi)W = W\Lambda$, where Λ is a diagonal eigenvalue matrix with its diagonal elements in decreasing order and W is an eigenvector matrix. However the matrix Q_Φ is always singular when S3 problem occurs. In this case, the traditional LDA method can not be used directly.

2.3 Two Parameters Regularization of Q_Φ

If all eigenvalues of Q_Φ are non-zero, the classical LDA method can be applied directly. In case of S3 problem occurs, LDA method can not be used since Q_Φ is singular. In designing the regularized matrix Q_Φ^R for the singular matrix Q_Φ, the criteria as suggested by Krzanowski etc [8] are used in this paper.

Assume $Q_\Phi = U_Q \Lambda_Q U_Q^T$ is the eigenvalue decomposition of matrix Q_Φ. Based on the results in [4], we define the two-parameter family regularization $Q_\Phi^{\alpha\beta}$ for Q_Φ as $Q_\Phi^{\alpha\beta} = U_Q \hat{\Lambda}_Q U_Q^T$, where $\hat{\Lambda}_Q$ is a diagonal matrix with its diagonal elements $\xi_i (i = 1, 2, ...d)$ given by,

$$\xi_i = \begin{cases} (\lambda_i + \alpha)/M \, , i = 1, 2, \cdots, \tau \\ \beta, \qquad\qquad i = \tau + 1, \cdots, N \end{cases} \tag{3}$$

where M is a normalization constant and is given by

$$M = \frac{tr(Q_\Phi) + \tau\alpha}{tr(Q_\Phi) - (N - \tau)\beta}, \tag{4}$$

where $\alpha \ge 0, \beta > 0$ and $(\xi_\tau + \alpha)/M - \beta \ge 0$. It is easily verified that the regularized matrix $Q_\Phi^{\alpha\beta}$ satisfies all the criteria listed in [8].

2.4 Formulating One Parameter Regularization

In this section, we derive the one parameter formulation from the above defined two parameters regularization.

Denote $G = diag(I_\tau, 0) \in R^{N \times N}$, $\bar{G} = I_N - G$, $a = w^T Q_\Phi w$, $b = w^T U_Q G U_Q^T w$, $c = w^T U_Q \bar{G} U_Q^T w$, $e = b + c = w^T w$. Then the regularized Fisher index $J_\Phi^{\alpha\beta}(w) = \frac{w^T P_\Phi w}{w^T Q_\Phi^{\alpha\beta} w}$, $w \in R^N$ can be written as

$$J_{\Phi}^{\alpha\beta}(w) = \frac{(tr(Q_{\Phi}) + \tau\alpha)\, w^T P_{\Phi} w}{(tr(Q_{\Phi}) - (N-\tau)\beta)\,(a + b\alpha) + c\beta\,(tr(Q_{\Phi}) + \tau\alpha)}.$$

Two optimal parameters α,β can be determined by solving equations $\nabla_{\alpha\beta} J_{\Phi}^{\alpha\beta}(w) = 0$, where ∇ is a gradient operator, as follows,

$$\alpha = \frac{tr(Q_{\Phi})c - (N-\tau)a}{bN - e\tau} \text{ and } \beta = \frac{tr(Q_{\Phi})}{N-\tau}.$$

On the other hand, we hope that above two parameters α, β can be reduced to one parameter t and when t tends to zero, the regularized matrix tends to the original matrix, i.e., $\alpha(t) \to 0$ and $\beta(t) \to 0$ as $t \to 0$. So we slightly modify the above formula as

$$\alpha(t) = \left| \frac{tr(Q_{\Phi})c - (N-\tau)a}{bN - e\tau} \right| \cdot t \text{ and } \beta(t) = \frac{tr(Q_{\Phi})}{N-\tau} \cdot t, \quad (0 < t < 1) \tag{5}$$

2.5 The Proposed 1PRKFD Algorithm

Based on results in sections 2.2 to 2.4, we develop one parameter regularized kernel Fisher discriminant (1PRKFD) algorithm for face recognition in this section. Details of the algorithm are designed as follows.

Step 1: Give initial value $\Theta = (\theta, t)$ and $w \in R^N$, calculate matrices Q_{Φ}, P_{Φ}.

Step 2: Do eigenvalue decomposition $Q_{\Phi} = U_Q \Lambda_Q U_Q^T$, where

$$\Lambda_Q = diag(\lambda_1, \cdots, \lambda_\tau, 0, \cdots, 0) \in R^{N \times N}, \lambda_1 > \lambda_2 > \cdots > \lambda_\tau > 0.$$

Step 3: Calculate α, β defined in (5) and ξ_i $(i = 1, 2, \cdots, N)$ defined in (3).

Step 4: Let $Y = \hat{\Lambda}_Q^{-1/2} U_Q^T$, $\hat{P}_{\Phi} = Y P_{\Phi} Y^T$, where $\hat{\Lambda}_Q = diag(\xi_1, \cdots, \xi_N)$, then do eigenvalue decomposition $\hat{P}_{\Phi} = V \Lambda_P V^T$, where Λ_P is a diagonal eigenvalue matrix of $\hat{P}_{\Phi}$ with its diagonal elements in decreasing order and V is an eigenvector matrix. Rewrite $V = (v_1, \cdots, v_{C-1}, \cdots, v_N)$ and let $V_{C-1} = (v_1, v_2, \cdots, v_{C-1})$.

Step 5: Calculate matrix $A_{K1PRDA} = U_Q \hat{\Lambda}_Q^{-1/2} V_{C-1}$. Rewrite $A_{K1PRDA} = [\tilde{w}_1, \tilde{w}_2, \cdots, \tilde{w}_{C1}]$, where $\tilde{w}_j = (\tilde{w}_{jk}^l)_{1 \le l \le C,\ 1 \le k \le N_j} \in R^N$,$1 \le j \le C-1$, and let $w_j = \sum_{l=1}^{C} \sum_{k=1}^{N_l} \tilde{w}_{jk}^l \Phi(x_k^l)$. Therefore the optimal projection matrix $W = [w_1, w_2, \cdots, w_{C-1}]$.

2.6 Determine the Optimal Parameters

In this section, the conjugate gradient method (CGM) will be exploited to determine the optimal parameters. The detail CGM algorithm is given as follows.

1. Give initial value $\Theta_1 = (\theta_1, t_1)$ and $w_0 \in R^N$, calculate matrices $Q_{\Phi}^{(1)}, P_{\Phi}^{(1)}$, via proposed 1PRKFD algorithm to get w_1.
2. Compute searching direction: $S_1 = \nabla J(\Theta_1, w_1)$, let $\hat{S}_1 = S_1 / \|S_1\|$
3. For $k \geq 1$**,** $\Theta_{k+1} = \Theta_k + \rho_k \cdot \hat{S}_k$**,** where $\hat{S}_k = S_k / \|S_k\|$, where $S_k = \nabla J(\Theta_k, w_k) + v_{k-1} \cdot S_{k-1}$ and

$$v_{k-1} = \|\nabla J(\Theta_k, w_k)\|^2 / \|\nabla J(\Theta_{k-1}, w_{k-1})\|^2 .$$

4. Calculate matrices $Q_{\Phi}^{(k+1)}, P_{\Phi}^{(k+1)}$, via 1PRKFD algorithm to obtain w_{k+1}. If $J(\Theta_{k+1}, w_{k+1}) < J(\Theta_k, w_k)$ then go to step 3 to search the next points.
5. The CGM iterative procedure of conjugate gradient method will terminate while $t < 0$ or $(\xi_\tau + \alpha)/M - \beta < 0$.

3 Experimental Results

To evaluate the proposed method, an in-depth investigation of the influence on performance of pose and illumination variations is performed using YaleB database. we select Gaussian RBF kernel as $K(x,y) = \exp(-2^{-1}\theta^2 \|x - y\|^2)$.

The YaleB database contains 5850 source images of 10 subjects each seen under 585 viewing conditions (9 poses × 65 illumination conditions). In our experiments, we use images under 45 illumination conditions and these images has been divided into four subsets according to the angle the light source direction makes with the camera axis [9].

3.1 Fixed Pose with Illumination Variations

In this part, we will investigate the influence of illumination variations upon performance of LDA-based face recognition algorithm. Results of fixing the pose and testing the illumination variations are shown in figure 1a-1b.

For each pose, we randomly select 2 images from each subset for training (2×4=8 images for training per individual), and all the other images from the 4 subsets are selected for testing (37 images for testing per individual). The experiments are repeated 10 times and the average accuracies of rank 1 are recorded and shown in the figure 1a. The mean accuracies of rank 1 to rank 3 of all poses are shown in figure 1b. In the CGM iterative procedure, the initial values of parameters are given as: $t = 0.005$, $\theta = 0.045$, the step length $\rho = 0.000125$ and $w_0 = ones(N, 1) \in R^N$.

From the results shown in Figure 1a, we can see that, (i) the proposed method gives comparable results with K1PRFD [6] and (ii) the performance of our method outperforms than other four methods under all illumination variations except that Kernel Direct LDA is slightly better than our method in pose 3.

From the results shown in Figure 1b, it can been seen that the recognition accuracy of our method increases from 88.92% (rank1) to 94.49%(rank3). The recognition accuracies of Eigenface [10], Fisherface [1], direct LDA [3], Kernel direct LDA [11] and K1PRFD [6] are increase from 58.98%, 70.42%, 80.39%,

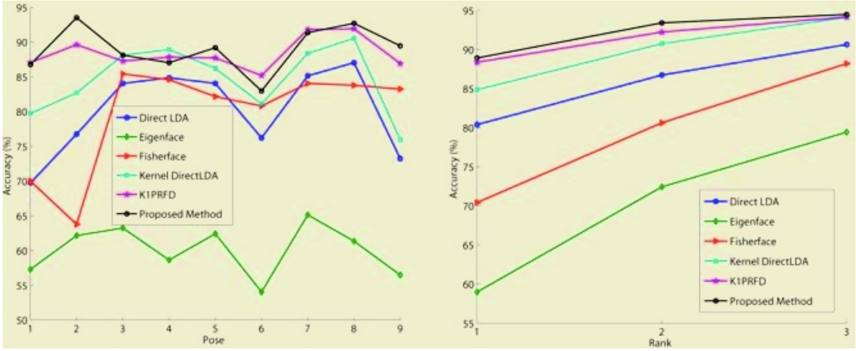


Fig. 1. For each pose, randomly select 2 images from each subset for training, and all the other images from the 4 subsets are selected for testing.

84.86% and 88.37% (rank1) to 79.43%, 88.2%, 90.63%, 94.11% and 94.17% (rank3) respectively. The results show that the proposed method outperforms than other five methods as well.

3.2 Both Pose and Illumination Variations

Finally, we will make the training samples include pose and illumination variations. The initial values of parameters are given as: $t = 0.005$, $\theta = 0.045$, the step length $\rho = 0.00125$ and $w_0 = ones(N, 1) \in R^N$. The experimental setting are as follows.

For each pose, we will select 2 images from each illumination subset of 4 subsets in all. This is to say that we will randomly select 720 images (10 persons × 9 poses × 4 subsets × 2 images) for training. Then the rest images, say 3330 images (10 persons × 9 poses × 37 images), are for testing. The experiments are repeated 10 times and the average rank1 to rank 3 accuracies are recorded and shown in the figure 2. From the results shown in Figure 2, it can be seen that the recognition accuracy of our method increases from 90.90% with rank 1 to 96.13% with rank 3. The results show that (i) the proposed method gives almost the same results with K1PRFD algorithm (the two curves are almost overlapped in Figure 2) and (ii) the proposed method outperforms than other five methods.

Finally, we would like to demonstrate the CGM iterative procedure. For fixed pose 2 with illumination variation case, the CGM starts from the initial values $\theta_0 = 0.0295$, $t_0 = 0.0039$, step length=0.00125 and $w_0 = ones(N, 1) \in R^N$. The CGM iterative procedure terminates at the iterative number 7, since the regularized parameter $t_7 = -0.0001 < 0$. The results show that the rank 1 accuracy increases from 86.49% with $\theta_1 = 0.029$, $t_1 = 0.0027$ to 90.27% with the final optimal parameter values $\theta_6 = 0.0373$, $t_6 = 0.0004$. The regularized parameter θ and the kernel parameter t versus Rank1 accuracy are recorded and plotted in the left side and right side of Figure 3 respectively.

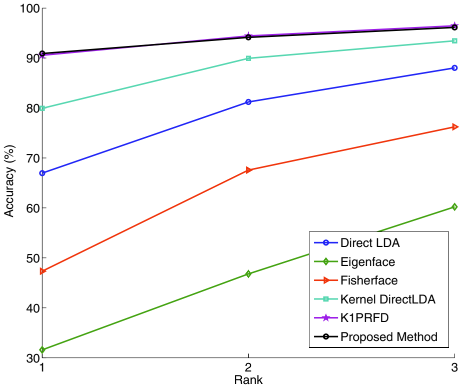


Fig. 2. Performance evaluation on pose and illumination.

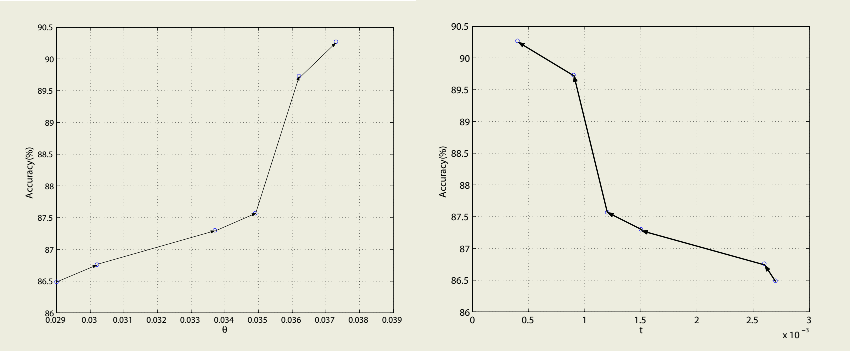


Fig. 3. Initial value $\theta_0 = 0.0295$, $t_0 = 0.0039$, step length=0.00125 and $w_0 = ones(N, 1) \in R^N$. The regularized parameter θ and the kernel parameter t versus Rank1 accuracy are recorded and plotted in the left side and right side of this figure respectively.

4 Conclusions

In this paper, a new one-parameter Regularization Kernel Fisher Discriminant (1PRKFD) is designed and developed based on two parameters regularized formula. We can select optimized regularized parameter t and kernel parameter θ of RBF kernel function simultaneously for 1PRKFD algorithm by performing conjugate gradient method (CGM). The results are encouraging on YeleB face databases. The performance are comparable with our previous developed K1PRFD method, but it outperforms with the existig LDA-based algorithm on pose and illumination variation. As the size of Yale B database is a relatively small, we will evaluate the proposed algorithm with larger databases in near future.

Acknowledgement

This project was supported by the Science Faculty Research grant of Hong Kong Baptist University RGC Earmarked Research Grant HKBU-2119/03E and NSFC (60144001, 10101013). The authors would like to thank for Yale University for contribution of the Yale Group B database.

References

1. P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.
2. S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K. R. Müller, "Fisher discriminant analysis with kernels," in *1999 IEEE Workshop on Neural Networks for Signal Processing IX*, 1999, pp. 41–48.
3. H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data - with application to face recogniion," *Pattern Recognition*, vol. 34, no. 10, pp. 2067–2070, 2001.
4. D. Q. Dai and P. C. Yuen, "Regularized discriminant analysis and its applications to face recognition," *Pattern Recognition*, vol. 36, no. 3, pp. 845– 847, 2003.
5. D. Q. Dai and P. C. Yuen, "A wavelet-based 2-parameter regularization discriminant analysis for face recognition", *Proceeding of The 4th International Conference on Audio and Video Based Personal Authentication*, June 2003
6. W. S. Chen, P C Yuen, J Huang and D. Q. Dai, "Kernel machine-based one-parameter regularized Fisher discriminant method for face recognition," *IEEE Transactions on SMC-B*, to appear, 2005.
7. S. Saitoh, *Theory of Reproducing Kernels and its Applications.* Harlow England: Longman Scientific Technical, 1988.
8. W. J. Krzanowski, P. Jonathan, W. V. McCarthy, and M. R. Thomas, "Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data," *Applied Statistics*, vol. 44, pp. 101–115, 1995.
9. A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.
10. M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.
11. J. W. Lu, K. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using kernel direct discriminant analysis algorithms," *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 117–126, Jan. 2003.

AutoAssign – An Automatic Assignment Tool for Independent Components

Matthias Böhm[1], Kurt Stadlthanner[1], Ana M. Tomé[2], Peter Gruber[1], Ana R. Teixeira[2], Fabian J. Theis[1], Carlos G. Puntonet[3], and Elmar W. Lang[1]

[1] Institute of Biophysics, University of Regensburg, D-93040 Regensburg, Germany
elmar.lang@biologie.uni-regensburg.de
[2] DETUA/IEETA, Universidade de Aveiro, P-3800 Aveiro, Portugal
ana@ieeta.pt
[3] Dept. Arquitectura y Tecnologia de Computadores, Universidad de Granada, E-18071 Granada, Spain

Abstract. In this work an automatic assignment tool for estimated independent components within an independent component analysis is presented. The algorithm is applied to the problem of removing the water artifact from 2D NOESY NMR spectra. The algorithm uses local PCA to approximate the water artifact and defines a suitable cost function which is optimized using simulated annealing. The blind source separation of the water artifact from the remaining protein spectrum is done with the recently developed algorithm dAMUSE.

1 Introduction

Blind Source Separation (BSS) methods consider the separation of observed sensor signals into their underlying source signals knowing neither these source signals nor the mixing process. Considering biomedical applications, BSS methods are especially valuable to remove artifacts from the signals recorded. In many biomedical applications quite a number of independent components have to be determined with ICA algorithms and it is not *a priori* clear how many components should be assigned to the signals representing the artifacts. This is especially obvious in 2D NOESY NMR proton spectra of proteins, where a prominent water artifact distorts the recorded spectra considerably. Recently artifact removal was considered using BSS techniques based on a generalized eigenvalue decomposition (GEVD) of a matrix pencil [5, 10]. Replacing the GEVD with the algorithm dAMUSE [8, 9], BSS and denoising can be achieved in one stroke. The method is very efficient and fast and outperformed FastICA and SOBI in all cases studied [7]. But, the estimated components related with the water artifacts had to be assigned by hand. With more than 100 estimated components this turns out to become a rather tedious undertaking prone to be biased by subjective judgements of the assignment criteria.

In this work we propose a local PCA approximation to the free induction decay (FID) related with the water artifact and to use simulated annealing [3] to

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 75–82, 2005.

determine those underlying uncorrelated components, estimated with dAMUSE, which have to be assigned to the water artifact.

The following section will provide a short summary of the algorithm dAMUSE [9] and introduces the new algorithm AutoAssign. To illustrate the proposed method, an application is discussed comprising theoretical 2D NOESY NMR spectra with added noise and an experimental water resonance added as well.

2 Theory

2.1 BSS Model

Given N complex sensor signals $x(t_{1,n}, t_{2,l}) \equiv x_n[l]$ sampled at L discrete time instances, they can be arranged in a data matrix $\mathbf{X}_{N \times L}$ with N rows and L columns, where the rows of the data matrix correspond to 1D free induction decays (FIDs) of the 2D NOESY experiment taken at N discrete evolution times $t_{1,n} \equiv [n], \; n = 1, \ldots, N$. Blind source separation (BSS) then relies on the following linear mixing model $\mathbf{x}[l] = \mathbf{A}\mathbf{s}[l] + \boldsymbol{\epsilon}[l]$ where $l = 0, \ldots, L-1$ and $\mathbf{x}[l] = (x_1[l], \ldots, x_N[l])^T$ designates the observed signals sampled at time instance l, $\mathbf{s}[l]$ the underlying uncorrelated source signals, $\mathbf{A}$ the stationary mixing matrix and $\boldsymbol{\epsilon}[l]$ an additional zero mean white Gaussian noise term which is independent of the source signals.

2.2 The Algorithm dAMUSE

A generalized eigenvalue decomposition using congruent matrix pencils may be used to separate water artifacts from 2D NOESY NMR spectra of proteins [6]. It provides the basis for the algorithm dAMUSE [9] used in the following, hence a short summary of the algorithm will be given.

Matrix Pencils: To solve the BSS problem we rely on second order GEVD techniques using congruent matrix pencils [10, 11]. First a matrix pencil $(\mathbf{R}_{x1}, \mathbf{R}_{x2})$ is computed with the sensor signals $\mathbf{x}[l]$, i.e. the observed FIDs. The $\mathbf{R}_{xj}, \; j = 1, 2$ denote corresponding correlation matrices of zero mean data. A GEVD of the sensor pencil then provides a solution for the BSS problem [10] and is given by $\mathbf{R}_{x1}\mathbf{E} = \mathbf{R}_{x2}\mathbf{E}\boldsymbol{\Lambda}$ where $\mathbf{E}$ represents a *unique* eigenvector matrix if the diagonal matrix $\boldsymbol{\Lambda}$ has *distinct* eigenvalues λ_i.

Embedding the Signals in Feature Space: Recently the GEVD using congruent matrix pencils has been extended to data embedded in a high-dimensional feature space of delayed coordinates to provide a means to perform BSS and denoising simultaneously [8, 9]. The method uses the concept of a trajectory matrix borrowed from singular spectral analysis (SSA) [1]. Consider a sensor signal component $x_n[l]$, each row of the trajectory matrix [4] contains delayed versions $x_n(l + (M - m)K)$, where K denotes the delay in number of sampling intervals between consecutive rows and M gives the dimension of the embedding space. Using a set of L samples and M delayed versions of the signal

$x_n[l + (M - m)K]$, $l = 0, \ldots, L-1$, $m = 0, \ldots, M-1$, the trajectory matrix is given by

$$(\mathbf{X}_n^e) = \begin{bmatrix} x_n[(M-1)K] & x_n[1+(M-1)K] & \cdots & x_n[L-1] \\ x_n[(M-2)K] & x_n[1+(M-2)K] & \cdots & x_n[L-1-K] \\ \vdots & \vdots & \ddots & \vdots \\ x_n[0] & x_n[1] & \cdots & x_n[L-1-(M-1)K] \end{bmatrix} \qquad (1)$$

The total trajectory matrix $\mathbf{X}^e$ of all N signals is formed by concatenating the component trajectory matrices $\mathbf{X}_n^e$ according to: $\mathbf{X}^e = [\mathbf{X}_1^e\mathbf{X}_2^e \ldots \mathbf{X}_N^e]^T$. After embedding, the instantaneous mixing model can be written as $\mathbf{X}^e = \mathbf{A}^e\mathbf{S}^e$ where $\mathbf{S}^e$ also represents the source signal trajectory matrix, $\mathbf{A}^e = \mathbf{A}_n \otimes \mathbf{I}_{M\times M}$ is a block matrix and $\mathbf{I}_{M\times M}$ denotes the identity matrix. Then if $\mathbf{A}_n$ is an invertible matrix, $\mathbf{A}^e$ is also invertible as it is the Kronecker product of two invertible matrices. The sensor pencil can be computed with $\mathbf{R}_{x1} = L^{-1}\mathbf{X}\mathbf{X}^H$ and $\mathbf{R}_{x2} = L^{-1}\mathbf{Z}\mathbf{Z}^H$ using the trajectory matrix $\mathbf{X}^e$ and a filtered version $\mathbf{Z}^e = \mathbf{X}\mathbf{C}^H$ with $\mathbf{C}$ a circular convolution matrix and H denoting the Hermitian conjugate [8]. The sensor pencil is again congruent with a corresponding source pencil, hence the respective eigenvectors are related by $\mathbf{E}_s^H = \mathbf{E}^H\mathbf{A}^e$. The linear transformation of the trajectory matrices then reads $\mathbf{Z}^e = \mathbf{E}^H\mathbf{X}^e = \mathbf{E}^H\mathbf{A}^e\mathbf{S}^e = \mathbf{E}_s^H\mathbf{S}^e$ Assuming that the source signals and their filtered versions are uncorrelated, the matrix $\mathbf{E}_s$ is block-diagonal, with block size $(M \times M)$.

Denoising Using the Algorithm dAMUSE: The eigenvalues and eigenvectors of a matrix pencil can be obtained via standard eigenvalue decompositions (EVD) applied in two consecutive steps. Considering the pencil $(\mathbf{R}_{x1}, \mathbf{R}_{x2})$ the following steps are performed:

- Compute a standard eigenvalue decomposition of the symmetric positive definite correlation matrix $\mathbf{R}_{x1} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^H$, i.e, the eigenvectors $(\mathbf{v}_i)$ and eigenvalues (λ_i) and organize the eigenvalues in descending order $(\lambda_1 > \lambda_2 > \ldots > \lambda_q \ldots > \lambda_{NM})$. For denoising purposes, a variance criterion has been established to retain only the largest eigenvalues exceeding a threshold parameter Θ [9].
- The transformation matrix can then be computed using the q largest eigenvalues and respective eigenvectors $Q = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{V}^H$ where $\mathbf{Q}$ is an $q \times NM$ matrix.
- Compute the matrix $\tilde{\mathbf{R}} = \mathbf{Q}\mathbf{R}_{x2}\mathbf{Q}^H$ and its standard eigenvalue decomposition: the eigenvector matrix $\mathbf{U}$ and eigenvalue matrix $\mathbf{D}_x$

The eigenvectors of the pencil $(\mathbf{R}_{x1}, \mathbf{R}_{x2})$ form the columns of the eigenvector matrix $\mathbf{E} = \mathbf{Q}^H\mathbf{U} = \mathbf{V}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}$ which can be used to compute the output signals as described above.

2.3 The Algorithm AutoAssign

Applying the BSS algorithms above to 2D NOESY NMR spectra to separate the water artifact and related artifacts from the protein spectra, the most tedious task is to assign the uncorrelated components estimated to the water signal. Because of erratic phase relations, up to 40 estimated components out of 128 or 256 need to be assigned to the water resonance. Hence an automated and objective assignment procedure deemed necessary.

The idea is to embed the signal in a high-dim feature space of delayed coordinates and to apply a cluster analysis to the columns of the corresponding trajectory matrix. Within each cluster a local PCA is then performed to obtain a low-dim approximation to the signals using only the most important principal components to approximate the signals. The latter are then feed into a suitable cost function which is optimized with simulated annealing.

Embedding and Local PCA: Consider a signal $\mathbf{x}_n[l] = (x_n[l], x_n[l+1], \dots, x_n[l+(M-1)])^T$ embedded in an M-dim feature space. Divide the space in k subspaces $\mathcal{N}^{(k)}$ using *k-means* clustering and center the signals in each cluster locally by subtracting the cluster mean $\bar{\mathbf{x}}_n^{(k)} = (\mathcal{N}^k)^{-1} \sum_{\mathbf{x}_n[l] \in \mathcal{N}^k} \mathbf{x}_n[l]$. Next a principal component analysis (PCA) is performed on each cluster separately. Then a local approximation $\mathbf{x}_{n,p}^{(k)}[l] = \sum_{j=1}^{p} \alpha_j[l]\mathbf{w}_j + \bar{\mathbf{x}}_n^{(k)}$ to the time domain signal is computed, using only the eigenvectors $\mathbf{w}_j$ to the p largest eigenvalues and $\alpha_j = \langle \mathbf{x}_{n,p}^{(k)}[l]\mathbf{w}_j \rangle$. This yields the new trajectory matrix $\mathbf{X}_{n,p}^{(k)}$ with entries $\mathbf{x}_{n,p}^{(k)}[l]$ and M delayed version thereof. The final local approximation $\langle \mathbf{x}_{n,p}^{(k)}[l] \rangle_{[l]}$ is obtained by averaging all entries at the same time instance (which lie along diagonals). Putting together all these local approximations yields the final approximation to the original signal observed.

As the water signal provides the dominant contribution to each FID observed, the approximation can be simplified further by retaining only the principal component to the largest eigenvalue, i.e. $\mathbf{x}_{n,1}[l] = \alpha_1[l]\mathbf{w}_1$. The approximation thus contains the contribution from the water signal almost exclusively.

Simulated Annealing: This approximation to the FID related with the water artifact is then used to define a cost function $\mathcal{E}(\boldsymbol{\beta}) = \sum_{l=0}^{L-1} (x_{n,\boldsymbol{\beta}}[l] - x_{n,1}[l])^2$ to be minimized with simulated annealing [3]. The BSS approximation to the water signal using the uncorrelated components estimated with the dAMUSE algorithm is obtained as $x_{n,\boldsymbol{\beta}}[l] = \sum_j \beta_j (\mathbf{A})_{nj} s_j[l]$ where a new configuration $\boldsymbol{\beta}$ is generated by changing any β_j randomly. A configuration is represented by a vector $\boldsymbol{\beta}$ which contains as many components β_j as there are sources $\mathbf{s}_j$. To each source one element of β is assigned which can take on the values $\beta_j \in \{0,1\}$ only. The difference in the values of the cost function for the current and the new configuration $\Delta\mathcal{E} = \mathcal{E}(\boldsymbol{\beta}_n) - \mathcal{E}(\boldsymbol{\beta}_{n+1})$ determines the probability of acceptance of the new configuration in the simulated annealing algorithm according to

$$\frac{P[\beta_{n+1}]}{P[\beta_n]} = \min\{1, \exp\left(-\frac{\Delta\mathcal{E}}{k_B T}\right)\} \tag{2}$$

After convergence, the configuration which best fits to the local PCA approximation of the water signal is obtained. Nullifying these components deliberately, the water-artifact-free protein spectrum $\tilde{\mathbf{x}}_n$ can be reconstructed using the remaining estimated source signals $\tilde{\mathbf{s}}_n$ via $\tilde{\mathbf{x}}_n = \mathbf{A}\tilde{\mathbf{s}}_n$.

3 Results and Discussion

The algorithms discussed above were applied to an artificial 2D NOESY proton NMR spectra of proteins dissolved in water. Every data set comprises 512 or 1024 FIDs $S(t_1, t_2) \equiv x_n[l]$, with $L = 2048$ samples each, which correspond to $N = 128$ or $N = 256$ evolution periods $t_1 \equiv [n]$. To each evolution period belong four FIDs with different phase modulations, hence only FIDs with equal phase modulations have been considered for analysis. A BSS analysis, using the algorithm dAMUSE, was applied to the FIDs collected in the data matrix $\mathbf{X}$. Filtering was done in the frequency domain for convenience. Hence, all FIDs have been Fourier transformed with respect to the sampling time t_2 to obtain 1D spectra $\hat{S}(t_1, \omega_2) \equiv \hat{\mathbf{x}}_n(\omega), \quad 0 < n \leq 128$ or $0 < n \leq 256$. The filtered versions of the data were obtained by applying a Gaussian filter $\hat{h}(\omega)$ with width $\sigma = 1$, centered near the water resonance, to each row of the data matrix $\hat{\mathbf{X}}$. After filtering the data have been back-transformed to the time domain to calculate the corresponding correlation matrices of the pencils. The automatic assignment of the uncorrelated components, estimated with dAMUSE, which belong to the water resonance was achieved using the proposed algorithm AutoAssign which is based on a local PCA and a simulated annealing optimization.

For test purposes, a theoretical 2D NOESY proton NMR spectrum of the cold-shock protein of the bacterium *Thematoga maritima*, containing only protein resonances, was used. The spectrum was obtained through reverse calculation using the algorithm RELAX [2]. Gaussian white noise with realistic amplitude as well as an experimentally recorded water resonance were added to the theoretical spectrum (see Fig. 1).

The BSS was done with the algorithm dAMUSE [9] using $K = 1$. The second correlation matrix $\mathbf{R}_{x2}$ of the pencil was computed with the theoretical spectrum with added noise taken as the filtered version of the original spectrum. An approximation of the water artifact dominating the time domain FIDs was obtained with the local PCA algorithm. As the experimental pure water FID was available also, both could be compared to access the quality of the approximation. To perform local PCA, each sample of the data was projected into a $M = 40$ dimensional feature space and *k-means* clustering was used to divide the projected data into $k = 2$ cluster. Only the largest principal component was considered to approximate the water signal.

Fig. 2-a) compares the total FID corresponding to the shortest evolution period with the local PCA approximation (Fig. 2-b)) of the latter. It is immediately obvious that the water artifact dominates the total FID. It is seen that local PCA provides a very good approximation to the water artifact. This can be corroborated by subtracting the approximate water FID from the total FID and transforming the resulting FID into the frequency domain. The resulting protein spectrum contains only small remnants of the huge water artifact as can be seen in Fig. 3. The spectra thus obtained will henceforth be called *approximated spectra*.

This indicates that it should be possible to use the local PCA approximation of the water artifact as a reference signal in a cost function to be min-

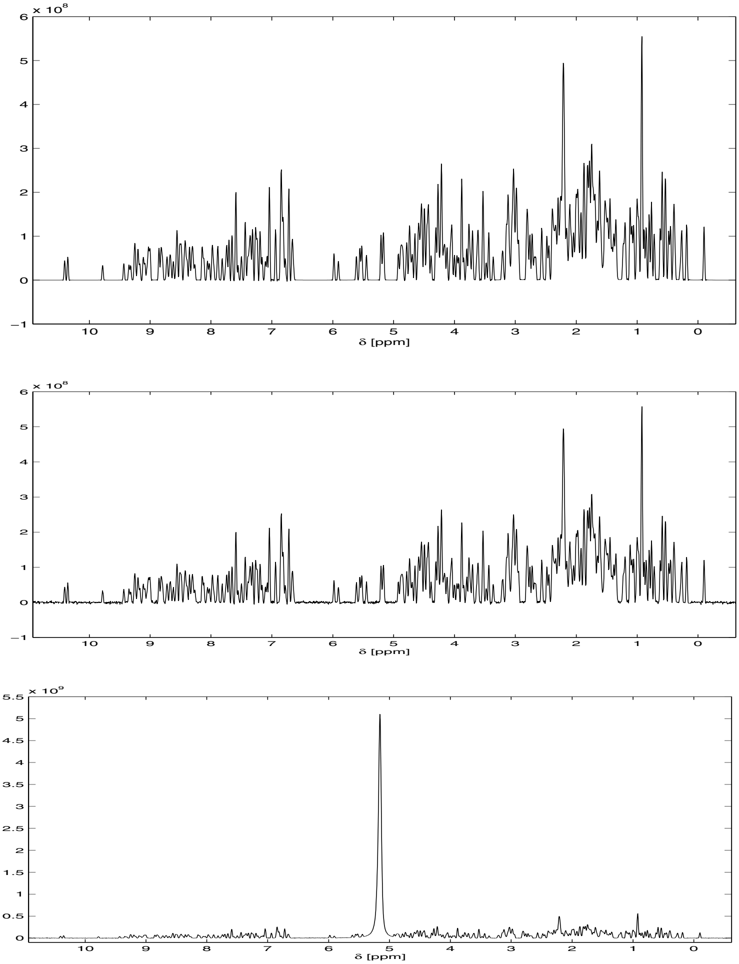


Fig. 1. (a) – Theoretical protein spectrum, (b) – Theoretical protein spectrum with Gaussian noise, (c) – Theoretical spectrum with Gaussian noise and an experimentally recorded water resonance.

imized with a simulated annealing algorithm. This is confirmed by analyzing the theoretical protein spectra plus noise plus water artifact (see Fig. 1-c)) with the dAMUSE algorithm to extract the uncorrelated components and using simulated annealing (SA) to automatically assign those components related with the water artifact. The SA-algorithm identifies the same 9 components irrespective whether the experimental water FID or its local PCA approximation has been used in the cost function. Without denoising, the reconstructed protein spectrum resembles the original noisy spectrum (Fig. 1-b)) except for a much enhanced noise level. Calculating the signal-to-noise ratio (SNR) via $SNR(\mathbf{x}, \mathbf{x}_{noise})[dB] = 20\log_{10}\frac{\|\mathbf{x}\|}{\|\mathbf{x}-\mathbf{x}_{noise}\|}$ where $\mathbf{x}$ denotes the theoretical spec-

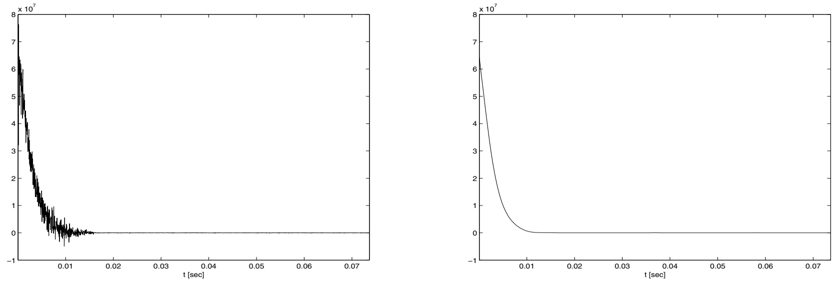


Fig. 2. Free Induction Decays (FID) of a) – the total FID of the protein signal plus additive noise plus an experimental water FID, b) – local PCA approximation of the total FID.

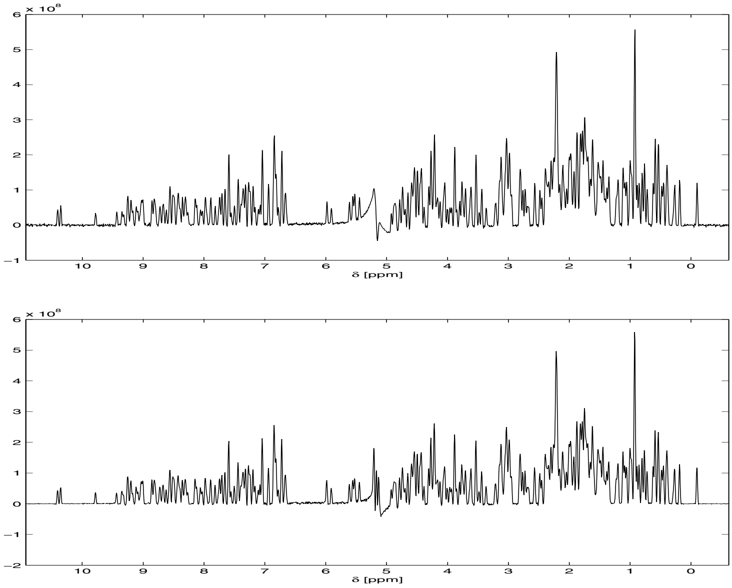


Fig. 3. a) – Protein spectrum obtained after subtracting the approximated water FID from the total FID and Fourier transformation of the difference FID, b) – Reconstructed protein spectrum obtained with dAMUSE.

trum, $\mathbf{x}_{noise}$ its noisy counterpart, the theoretical spectrum plus gaussian noise shows a SNR of $24.5dB$, whereas the reconstructed protein spectrum only yields a SNR of $10.1dB$. Denoising can be accomplished elegantly with the dAMUSE algorithm which achieves blind source separation and denoising simultaneously. The water related components extracted are automatically assigned with the algorithm AutoAssign using a local PCA approximation to the water artifact. The optimal number M of delays as well as the optimal size of the time lag have been determined by the best minimum to the cost function obtained with the SA algorithm. A minimum of the cost function has been obtained with using one time-delayed FID, a lag of one sampling interval and by retaining 158 eigenvectors (out of $2 \cdot 128$) after the first step of the algorithm dAMUSE. The result of the dAMUSE denoising is shown in Fig. 3, the SNR achieved amounts to $SNR = 22.1\ dB$.

References

1. M. Ghil, M.R. Allen, M. D. Dettinger, K. Ide, and et al. Advanced spectral methods for climatic time series. *Reviews of Geophysics*, 40(1):3.1–3.41, 2002.
2. A. Görler and H. R. Kalbitzer. RELAX, a flexible program for the back calculation of NOESY spectra based on a complete relaxation matrix formalism. *Journal of Magnatic Resonance*, 124:177–188, 1997.
3. S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:pp. 671, 1983.
4. V. Moskvina and K. M. Schmidt. Approximate projectors in singular spectrum analysis. *SIAM Journal Mat. Anal. Appl.*, 24(4):932–942, 2003.
5. K. Stadlthanner, F. Theis, E. W. Lang, A. M. Tomé, W. Gronwald, and H. R. Kalbitzer. A matrix pencil approach to the blind source separation of artifacts in 2D NMR spectra. *Neural Information Processing - Letters and Reviews*, 1:103 – 110, 2003.
6. K. Stadlthanner, A. M. Tomé, F. J. Theis, W. Gronwald, H. R. Kalbitzer, and E. W. Lang. Blind source separation of water artifacts in NMR spectra using a matrix pencil. In *Fourth International Symposium On Independent Component Analysis and Blind Source Separation, ICA'2003*, pages 167–172, Nara, Japan, 2003.
7. K. Stadlthanner, A. M. Tomé, F. J. Theis, W. Gronwald, H. R. Kalbitzer, and E. W. Lang. On the use of independent component analysis to remove water artifacts of 2D NMR protein spectra. In *7th Portuguese Conference on Biomedical Engineering, BIOENG'2003*, Lisbon, Portugal, 2003.
8. A. R. Teixeira, A. M. Tomé, E. W. Lang, and K. Stadlthanner. Delayed AMUSE - A Tool for blind source separation and Denoising. In *Independent Component Analysis and Blind Signal Separation, Proc. ICA'2004*, volume LNCS 195, pages 287–294, 2004.
9. Ana Maria Tomé, Ana Rita Teixeira, Elmar Wolfgang Lang, Kurt Stadlthanner, and A.P. Rocha. Blind source separation using time-delayed signals. In *International Joint Conference on Neural Networks, IJCNN'2004*, volume CD, Budapest, Hungary, 2004.
10. Ana Maria Tomé. An iterative eigendecomposition approach to blind source separation. In *3rd Intern. Conf. on Independent Component Analysis and Signal Separation, ICA'2003*, pages 424–428, San Diego, USA, 2001.
11. Lang Tong, Ruey-wen Liu, Victor C. Soon, and Yih-Fang Huang. Indeterminacy and identifiability of blind identification. *IEEE Transactions on Circuits and Systems*, 38(5):499–509, 1991.

Improving the Discrimination Capability with an Adaptive Synthetic Discriminant Function Filter

J. Ángel González-Fraga[1], Víctor H. Díaz-Ramírez[1],
Vitaly Kober[1], and Josué Álvarez-Borrego[2]

[1] Department of Computer Sciences, Division of Applied Physics
CICESE, Ensenada, B.C. 22860, Mexico
vkober@cicese.mx
[2] Optics Department, Division of Applied Physics
CICESE, Ensenada, B.C. 22860, Mexico
josue@cicese.mx

Abstract. In this paper a new adaptive correlation filter based on synthetic discriminant functions (SDF) for reliable pattern recognition is proposed. The information about an object to be recognized and false objects as well as background to be rejected is used in an iterative procedure to design the adaptive correlation filter with a given discrimination capability. Computer simulation results obtained with the proposed filter in test scenes are compared with those of various correlation filters in terms of discrimination capability.

1 Introduction

Since the introduction of the matched spatial filter (MSF) [1], many different types of filters for pattern recognition based on correlation have been proposed [2-11]. The traditional way to design correlation filters is to make filters that optimize different criteria. Several performance measures for correlation filters have been proposed and summarized in [5]. Some of the measures can be essentially improved using an adaptive approach to the filter design. According to this concept we are interested in a filter with good performance characteristics for a given observed scene, i.e. with a fixed set of patterns or a fixed background to be rejected, rather than to construct a filter with average performance parameters over an ensemble of images.

One of the most important performance criteria in tasks of pattern recognition is the discrimination capability, or how well a filter detects and discriminates between classes of objects. A theoretical analysis of correlation methods has been made by Yaroslavsky [6]. He suggested a filter with minimum probability of anomalous localization errors (false alarms) and called the optimal filter (OF). An important feature of the OF is its adaptivity in application to pattern recognition or target detection because its frequency response takes into account the power spectrum of wrong objects or an observed scene background to be rejected. A disadvantage of the OF in optical implementation is its extremely low light efficiency. The correlation filter with maximal light efficiency is a phase-only filter (POF) [2]. An approximation of

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 83–90, 2005

the OF by means of phase-only filters with a quantization was made in [7]. There approximate filters with high light efficiency and discrimination capability close to that of the OF were proposed and investigated. Another fruitful approach to the synthesis of adaptive filters with improved capability to discriminate between similar objects was proposed in [10].

An attractive approach to distortion-invariant pattern recognition is the use of a synthetic discriminant function (SDF) filter [3, 4]. The SDF filters use a set of training images to synthesize a template that yields a prespecified central correlation outputs in response to training images. The main shortcoming of the SDF filters is appearance of sidelobes owing to the lack of control over the whole correlation plane in the SDF approach. As a result, the SDF filters often possess a low discrimination capability. A partial solution to this problem was suggested [11].

In this work, a new adaptive SDF filter algorithm is proposed for elimination of sidelobes and improving the discrimination capability. The proposed filter is designed to reject sidelobes of an input scene background as well as false objects. In such a way we are able to control the whole correlation plane. The performance of the adaptive filter in test scenes are compared with those of the MSF, the POF and the OF in terms of discrimination capability.

Section 2 is a review of SDF filters. The design of the adaptive SDF algorithm is given in Section 3. Computer simulation results are presented and discussed in Section 4. Finally, conclusions of our work are provided in Section 5.

2 Synthetic Discriminant Functions

The SDF filter is a linear combination of MSFs for different patterns [3, 4]. The coefficients of the linear combination are chosen to satisfy a set of constraints on the filter output, requiring a prespecified value for each of the patterns used in the filter synthesis

2.1 Intraclass Recognition Problem

Let $\{f_i(x,y)\}$, $i=1,2,...,N$ be a set of (linearly independent) training images, each with d pixels. The SDF filter function $h(x,y)$ can be expressed as a linear combination of the set of reference images $f_i(x,y)$, i.e.

$$h(x,y)=\sum_{i=1}^{N} a_i f_i(x,y) \tag{1}$$

where a_i are weighting coefficients, which are chosen to satisfy the following conditions:

$$f_i \circ h = u_i \tag{2}$$

where the symbol $\circ$ denotes correlation and u_i is a prespecified value in the correlation output at the origin for each training image.

Let $\mathbf{R}$ denote a matrix with N columns and d rows (number of pixels in each training image), where its *i*th column is given by the vector version of $f_i(x, y)$. Let $\mathbf{a}$ and $\mathbf{u}$ represent column vectors of the elements a_i and u_i, respectively. We can rewrite equations (1) and (2) in matrix-vector notation as follows:

$$\mathbf{h} = \mathbf{aR}, \tag{3}$$

$$\mathbf{u} = \mathbf{R}^{+}\mathbf{h}, \tag{4}$$

where superscript + means conjugate transpose. By substituting equation (3) into equation (4) we obtain

$$\mathbf{u} = (\mathbf{R}^{+}\mathbf{R})\mathbf{a}. \tag{5}$$

The element (i, j) of the matrix $\mathbf{S} = (\mathbf{R}^{+}\mathbf{R})$ is the value at the origin of the cross-correlation between the training images $f_i(x, y)$ and $f_j(x, y)$. If the matrix $\mathbf{S}$ is non-singular, the solution of the equation system is

$$\mathbf{a} = (\mathbf{R}^{+}\mathbf{R})^{-1}\mathbf{u}, \tag{6}$$

and the filter vector is given by

$$\mathbf{h} = \mathbf{R}(\mathbf{R}^{+}\mathbf{R})^{-1}\mathbf{u}. \tag{7}$$

An equal correlation peak SDF filter can be used for intraclass distortion-invariant pattern recognition, (i.e., recognition of several images obtained from the true class objects). This can be done by setting all elements of $\mathbf{u}$ to unity, i.e.

$$\mathbf{u} = [11...1]^{T}. \tag{8}$$

2.2 Multiclass Recognition Problem

Now assume that there are a distorted version of the reference and various other classes of objects to be rejected. For simplicity, we consider two-class recognition problem. Thus, we design a filter to recognize training images from one class (called the true class) and to reject training images from another class (called the false class).

Suppose that there are M training images for the false class $\{p_i(x, y)\}, i = 1,2,...,M$. According to the SDF approach, the composite image $h(x, y)$ is a linear combination of the training images; that is, $\{f_1(x, y),..., f_N(x, y), p_1(x, y),..., p_M(x, y)\}$. The both intraclass recognition and interclass discrimination (i.e., discrimination of the true class objects against the false class objects) problems can be solved by means of SDF filters. We can set $u_i = 1$ for the true class objects and $u_i = 0$ for the false class objects as follows:

$$\mathbf{u} = [11...100...0]^{T} \tag{9}$$

Using the filter in (7) for pattern recognition, we expect that the central correlation will be close to 1 for the true class objects and it will be close to 0 for the false class objects. Obviously the above approach can be extended to any number of classes (in theory). Unfortunately, this simple procedure is the lack of control over the whole correlation output. This means that sidelobes may appeared anywhere on the correlation plane.

3 Design of Adaptive SDF Filter

A new adaptive SDF filter is proposed to recognize objects in high cluttered input scenes. We use discrimination capability (DC) for the filter design. The DC for pattern recognition is defined [5, 10] as the ability of a filter to distinguish a target among other different objects. If a target is embedded into a background that contains false objects, then the DC can be expressed as follows:

$$DC = 1 - \frac{\left|C^B(0,0)\right|^2}{\left|C^T(0,0)\right|^2} \tag{10}$$

where C^B is the maximum in the correlation plane over the background area to be rejected, and C^T is the maximum in the correlation plane over the area of object to be recognized. The area of the object to be recognized is determined in the close vicinity of the target location (the size of the area is similar to the size of the target). The background area is complementary to the object area. Negative values of the DC indicate that a tested filter fails to recognize the target.

We are interested in a filter that identifies the target with a high discrimination capability in high cluttered and noisy input scenes. In this case, actually conventional correlation filters yield a poor performance because of a low tolerance to noise and false details. With help of the adaptive SDF filters a given value of the DC can be achieved.

The algorithm of the filter design requires knowledge of the background image. This means that we are looking the target with unknown location in the known input scene background. The background can be described either stochastically, for instance, it can be considered as a realization of stochastic process or deterministically, that can be a given picture. The first step is to carry out correlation between the background and a basic filter SDF filter, which is trained only with the target. Next, the maximum of the filter output is set as the origin for the next iteration of training. Now two-class recognition problem is utilized to design the SDF filter; that is, the true class is the target and the false class is the background with the region of support equals to that of the target. The described iterative procedure is carried out while a given value of the DC is obtained. A block-diagram of the procedure is shown in Fig. 1.

Finally, note that if other objects to be rejected are known, they can be directly included in the SDF filter design.

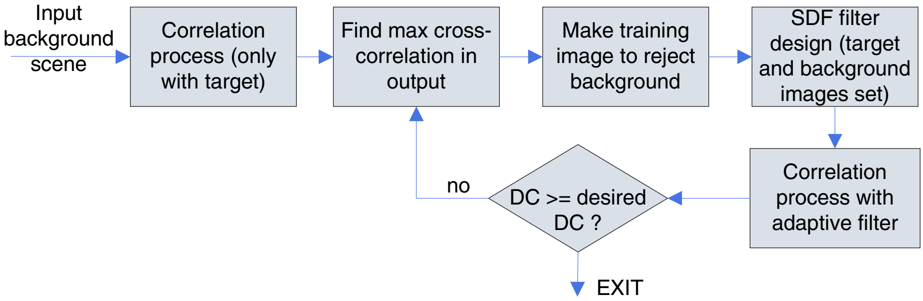
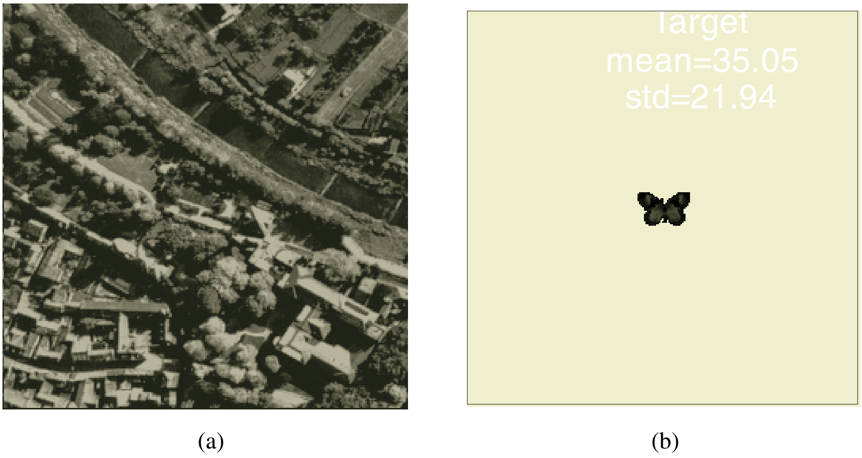

Input background scene
Correlation process (only with target)
Find max cross-correlation in output
Make training image to reject background
SDF filter design (target and background images set)
Correlation process with adaptive filter
DC >= desired DC ?
no
EXIT

Fig. 1. Block diagram of the iterative process to design the adaptive SDF filter.

4 Computer Simulation

In this section simulation results obtained with the adaptive SDF filter are presented. These results are compared with those of the MSF, the POF and the OF filters. The size of all images used in our experiments is 256×256 pixels. The signal range is [0-255]. The size of the butterflies is about 30×20 pixels. We use a real background shown in Fig. 2 (a) as an input scene. The average and standard deviation of the background are 84 and 40, respectively. The target is a butterfly shown in Fig. 2(b). The average and the standard deviation over the target area are 35 and 22, respectively. The signal to noise ratio of these signals is 0.0017.



(a)

(b)

Fig. 2. (a) Real background used in experiments, (b) target.

Figure 3 shows the performance of the adaptive filter in terms of the DC versus the iteration number. One can observe that before iteration 2 the DC is negative. After 20 iterations the obtained adaptive filter yields DC=0.998. All calculations are made with real values. This means that a high level of the correlation plane control can be achieved for a given input scene.

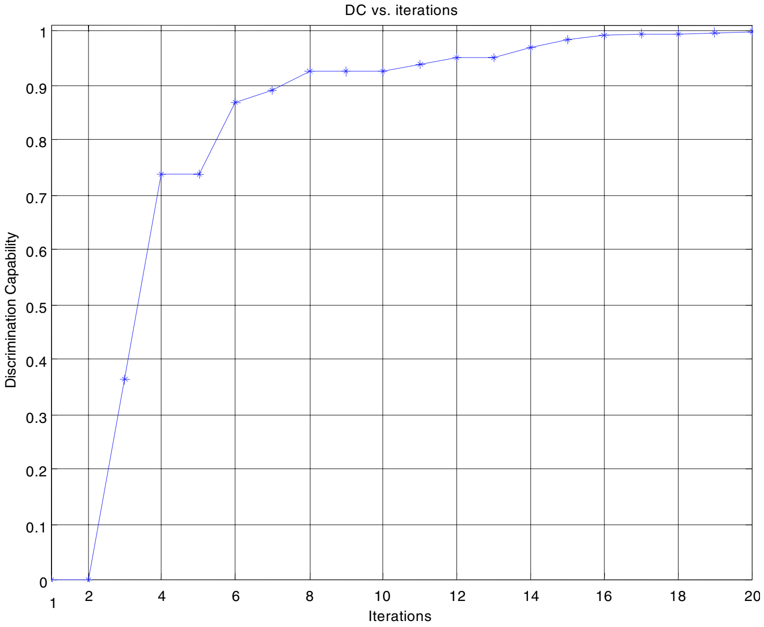


Fig. 3. Performance of the adaptive SDF filter at each iteration.

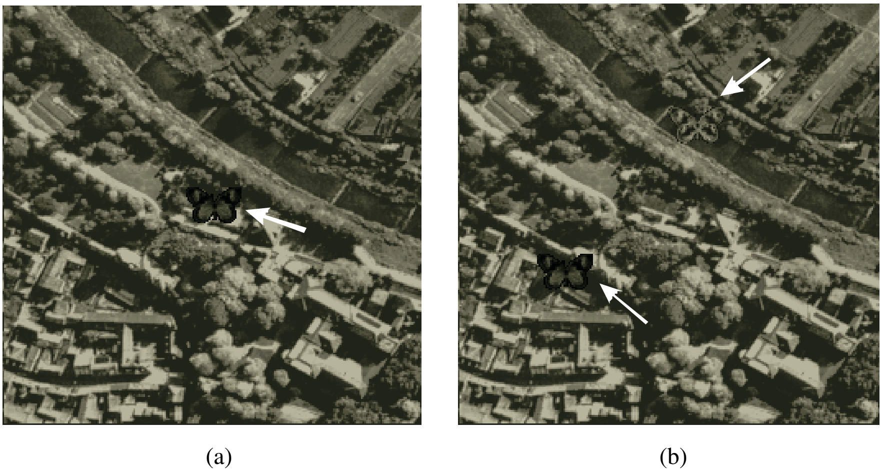


Fig. 4. Test scenes. (a) target is marked with a white arrow, (b) target is in the lower left corner and false butterfly in the upper right corner.

Next, we test the performance of pattern recognition with the adaptive filter when the target is placed into the background at arbitrary coordinates. The input scene is shown in Fig. 4 (a). The performance of the MSF, the POF and the OF in terms of the DC is given in line 1 of Table 1. Obviously, the proposed filter referred to as A-SDF1 gives the best performance in terms of the DC. We used 30 statistical trials of our

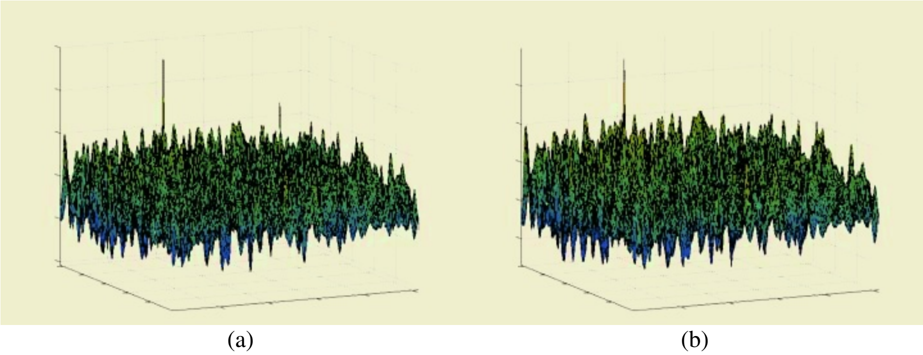
Table 1. Performance of different correlation filters in terms of DC.

Scene	MSF	POF	OF	A-SDF1	A-SDF2
a	-0.988	-0.422	0.626	0.992	-
b	-2.507	-2.995	-1.374	0.633	0.985

experiment for different positions of the target. With 95% confidence the DC is equal to 0.992±0.007.

Next, we place a false butterfly into the background. The average and standard deviation over the false object are 74 and 29, respectively. This scene is shown in Fig. 4(b). The adaptive filter design was made taking into account the false butterfly. We called the second adaptive filter as A-SDF2. Pattern recognition with the adaptive filter when the target is embedded into the background at arbitrary coordinates was performed. The input scene is shown in Fig. 4 (b). The performance of the correlation filters in terms of the DC is given in line 2 of Table 1. In this case the proposed adaptive yields also the best performance in terms of the DC. With 95% confidence the DC is equal to 0.985±0.009.

Figure 5 shows the correlation planes obtained with A-SDF1 and A-SDF2 filters for two test scenes; that is, Fig. 5(a) is the filter output with A-SDF1 for Fig. 4 (a) and Fig. 5(b) is the filter output with A-SDF2 for Fig. 4 (b).


(a)

(b)

Fig. 5. Correlation distributions obtained with (a) adaptive SDF filter A-SDF1 for the test scene in Fig. 4(a), (b) adaptive SDF filter A-SDF2 for the test scene in Fig. 4(b).

5 Conclusion

In this paper, new adaptive SDF filters have been proposed to improve recognition of a target embedded into a known cluttered background. We compared the performance of pattern recognition with various popular correlation filters and the proposed adaptive SDF filters in terms of discrimination capability. The computer simulation results have shown the superiority of the proposed filters comparing with the MSF, the POF, and the OF filters.

References

1. Vander Lugt A. B., Signal detection by complex filtering. IEEE Trans. Inf. Theory, Vol. 10, (1964) 139-135.
2. Horner J. L. and Gianino P. D., Phase-only matched filtering. Applied Optics, Vol. 23, (1984) 812-816.
3. Casasent. D., Unified synthetic discriminant function computational formulation. Applied Optics, Vol. 23, (1984) 1620-1627.
4. Vijaya Kumar B. V. K., Tutorial survey of composite filter designs for optical correlators. Applied Optics, Vol. 31, No. 23, (1992) 4773-4801.
5. Vijaya Kumar B. V. K. and Hassebrook L., Performance measures for correlation filters. Applied Optics, Vol. 29, No. 20, (1990) 2997-3006.
6. Yaroslavsky L. P., The theory of optimal methods for localization of objects in pictures. in progress in Optics XXXII, E. Wolf, Ed., Elsevier, (1993) 145-201
7. Kober V., Yaroslavsky L.P. Campos J., and Yzuel M.J., Optimal filter approximation by means of a phase only filter with quantization. Optics Letters, Vol. 19, No. 13, (1994) 978-980.
8. Javidi B. and Wang J., Design of filters to detect a noisy target in nonoverlapping background noise. Journal OSA (A), Vol. 11, (1994) 2604-2612.
9. Kober V. and Campos J., Accuracy of location measurement of a noisy target in a nonoverlapping background. Journal OSA (A), Vol. 13, (1996)1653-1666.
10. Kober V. and Ovseyevich I.A., Phase-only filter with improved filter efficiency and correlation discrimination. Pattern Recognition and Image Analysis, Vol. 10, No 4, (2000) 514-519.
11. Mahalanobis A., Vijaya Kumar B. V. K., Casasent. D., Minimum average correlation filters. Applied Optics, Vol. 26, (1987) 3633-3640.

Globally Exponential Stability of Non-autonomous Delayed Neural Networks*

Qiang Zhang[1], Wenbing Liu[2], Xiaopeng Wei[1], and Jin Xu[1]

[1] University Key Lab of Information Science & Engineering, Dalian University,
Dalian, 116622, China
zhangq30@yahoo.com

[2] School of Computer Science and Engineering, Wenzhou Normal College,
Wenzhou, 325027, China

Abstract. Globally exponential stability of non-autonomous delayed neural networks is considered in this paper. By utilizing delay differential inequalities, a new sufficient condition ensuring globally exponential stability for non-autonomous delayed neural networks is presented. The condition does not require that the delay function be differentiable or the coefficients be bounded. Due to this reason, the condition improves and extends those given in the previous literature.

1 Introduction

Autonomous delayed neural networks(DNNs) have been extensively studied in the past decade and successfully applied to signal-processing systems, static image treatment, patter recognition, associative memories and to solve nonlinear algebraic equations. Such applications rely on qualitative properties of stability. For this reason, the stability of autonomous delayed neural networks have been deeply studied and many important results on the global asymptotic stability and global exponential stability of one unique equilibrium point have been presented, see, for example,[1]-[21] and references cited therein. However, to the best of our knowledge, few studies have considered dynamics for non-autonomous delayed neural networks [22]. In this paper, by using a delay differential inequality, we discuss the globally exponential stability of non-autonomous delayed neural networks and obtain a new sufficient condition. We do not require the delay to be differentiable.

2 Preliminaries

The dynamic behavior of a continuous time non-autonomous delayed neural networks can be described by the following state equations:

$$\begin{aligned} x_i'(t) &= -c_i(t)x_i(t) + \sum_{j=1}^{n} a_{ij}(t) f_j(x_j(t)) \\ &\quad + \sum_{j=1}^{n} b_{ij}(t) f_j(x_j(t-\tau_j(t))) + I_i(t). \end{aligned} \tag{1}$$

* The project was supported by the National Natural Science Foundation of China (Grant Nos. 60403001, 60403002)and China Postdoctoral Science Foundation.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 91–96, 2005.

where n corresponds to the number of units in a neural networks; $x_i(t)$ corresponds to the state vector at time t; $f(x(t)) = [f_1(x_1(t)), \cdots, f_n(x_n(t))]^T \in R^n$ denotes the activation function of the neurons; $A(t) = [a_{ij}(t)]_{n \times n}$ is referred to as the feedback matrix, $B(t) = [b_{ij}(t)]_{n \times n}$ represents the delayed feedback matrix, while $I_i(t)$ is a external bias vector at time t, $\tau_j(t)$ is the transmission delay along the axon of the jth unit and satisfies $0 \leq \tau_i(t) \leq \tau$.

Throughout this paper, we will assume that the real valued functions $c_i(t) > 0, a_{ij}(t), b_{ij}(t), I_i(t)$ are continuous functions. The activation functions $f_i, i = 1, 2, \cdots, n$ are assumed to satisfy the following conditions (H)

$$|f_i(\xi_1) - f_i(\xi_2)| \leq L_i|\xi_1 - \xi_2| \ , \forall \xi_1, \xi_2 .$$

This type of activation functions is clearly more general than both the usual sigmoid activation functions and the piecewise linear function (PWL): $f_i(x) = \frac{1}{2}(|x+1| - |x-1|)$ which is used in [5].

The initial conditions associated with system (1) are of the form

$$x_i(s) = \phi_i(s), \ s \in [-\tau, 0], \ \tau = \max_{1 \leq i \leq n} \{\tau_i^+\}$$

in which $\phi_i(s)$ are continuous for $s \in [-\tau, 0]$.

Throughout this paper, we denote D^+ as the upper right Dini derivative. For any continuous function $f : R \to R$, the upper right Dini derivative of $f(t)$ is defined as

$$D^+ f(t) = \lim_{\delta \to 0^+} \sup \frac{f(t+\delta) - f(t)}{\delta}$$

Lemma 1. *[23] Let $x(t)$ be a continuous nonnegative function on $t \geq t_0 - \tau$ satisfying inequality (2) for $t \geq t_0$.*

$$D^+ x(t) \leq -k_1(t) x(t) + k_2(t) \bar{x}(t) \tag{2}$$

where $\bar{x}(t) = \sup\limits_{t-\tau \leq s \leq t} \{x(s)\}$. If $k_1(t)$ or $k_2(t)$ is bounded, and $\alpha = inf_{\{t \geq t_0\}}$ $\{k_1(t) - k_2(t)\} > 0$, then there must exist a positive $\eta > 0$ such that

$$x(t) \leq \bar{x}(t_0) \exp\{-\eta(t - t_0)\} \tag{3}$$

holds for all $t \geq t_0 - \tau$.

3 Global Exponential Stability Analysis

In this section, we will use the above Lemma to establish the exponential stability of system (1). Consider two solutions $x(t)$ and $z(t)$ of system (1) for $t > 0$ corresponding to arbitrary initial values $x(s) = \phi(s)$ and $z(s) = \varphi(s)$ for $s \in [-\tau, 0]$. Let $y_i(t) = x_i(t) - z_i(t)$, then we have

$$\begin{aligned} y_i'(t) = -c_i(t) y_i(t) &+ \sum_{j=1}^{n} a_{ij}(t) \left(f_j(x_j(t)) - f_j(z_j(t)) \right) \\ &+ \sum_{j=1}^{n} b_{ij}(t) \left(f_j(x_j(t - \tau_j(t))) - f_j(z_j(t - \tau_j(t))) \right) \end{aligned} \tag{4}$$

Set $g_j(y_j(t)) = f_j(y_j(t) + z_j(t)) - f_j(z_j(t))$, one can rewrite Eq.(4) as

$$y_i'(t) = -c_i(t)y_i(t) + \sum_{j=1}^{n} a_{ij}(t)g_j(y_j(t)) + \sum_{j=1}^{n} b_{ij}(t)g_j(y_j(t - \tau_j(t))) \tag{5}$$

Note that the functions f_j satisfy the hypothesis (H), that is,

$$\begin{gathered} |g_i(\xi_1) - g_i(\xi_2)| \leq L_i|\xi_1 - \xi_2| \ , \forall \xi_1, \xi_2. \\ g_i(0) = 0 \end{gathered} \tag{6}$$

From Eq.(5), we can get

$$D^+|y_i(t)| \leq -c_i(t)|y_i(t)| + \sum_{j=1}^{n} L_j|a_{ij}(t)||y_j(t)| + \sum_{j=1}^{n} L_j|b_{ij}(t)||\bar{y}_j(t)| \tag{7}$$

Theorem 1. *Let*

$$\begin{aligned} k_1(t) &= \min\nolimits_i \left[c_i(t) - \sum_{j=1}^{n} \tfrac{\alpha_j}{\alpha_i} L_i |a_{ji}(t)| \right] \\ k_2(t) &= \max\nolimits_i \left(\sum_{j=1}^{n} \tfrac{\alpha_j}{\alpha_i} L_i |b_{ji}(t)| \right) \end{aligned} \tag{8}$$

where $\alpha_i > 0$ is a positive constant. Eq.(1) is globally exponentially stable if

$$\alpha = \inf_{t \geq t_0} \{k_1(t) - k_2(t)\} > 0$$

Proof. Let $z(t) = \sum_{i=1}^{n} \alpha_i |y_i(t)|$, Calculating the Dini derivative of $z(t)$ along the solutions of (5), we get

$$\begin{aligned} D^+z(t) &\leq \sum_{i=1}^{n} \alpha_i D^+|y_i(t)| \\ &\leq \sum_{i=1}^{n} \alpha_i \left[-c_i(t)|y_i(t)| + \sum_{j=1}^{n} L_j|a_{ij}(t)||y_j(t)| \right. \\ &\qquad \left. + \sum_{j=1}^{n} L_j|b_{ij}(t)||\bar{y}_j(t)| \right] \\ &= -\sum_{i=1}^{n} \alpha_i \left[c_i(t) - \sum_{j=1}^{n} \frac{\alpha_j}{\alpha_i} L_i|a_{ji}(t)| \right] |y_i(t)| \\ &\quad + \sum_{i=1}^{n} \alpha_i \left[\sum_{j=1}^{n} \frac{\alpha_j}{\alpha_i} L_i|b_{ji}(t)| \right] |\bar{y}_i(t)| \\ &\leq -k_1(t)z(t) + k_2(t)\bar{z}(t) \end{aligned}$$

According to Lemma above, if the condition (3) is satisfied, then we have

$$\begin{aligned}\alpha_{\min}\sum_{i=1}^{n}|y_i(t)| &\leq z(t)=\sum_{i=1}^{n}\alpha_i|y_i(t)|\\ &\leq \bar{z}(t_0)\exp\{-\eta(t-t_0)\}\\ &=\sum_{i=1}^{n}\alpha_i|\bar{y}_i(t_0)|\exp\{-\eta(t-t_0)\}\\ &\leq \alpha_{\max}\sum_{i=1}^{n}|\bar{y}_i(t_0)|\exp\{-\eta(t-t_0)\}\end{aligned}$$

which implies that $\sum_{i=1}^{n}|y_i(t)| \leq \frac{\alpha_{\max}}{\alpha_{\min}}\sum_{i=1}^{n}|\bar{y}_i(t_0)|\exp\{-\eta(t-t_0)\}$. This completes the proof.

Remark 1. Note that the criteria obtained here are independent of delay and the coefficients $c_i(t), a_{ij}(t)$ and $b_{ij}(t)$ may be unbounded.

4 An Illustrative Example

In this section, we will give an example showing the effectiveness of the condition given here.

Example 1. Consider the following non-autonomous delayed neural networks

$$\begin{aligned}x_1^{'}(t) &= -c_1(t)x_1(t)+a_{11}(t)f(x_1(t))+a_{12}(t)f(x_2(t))\\ &\quad +b_{11}(t)f(x_1(t-\tau_1(t)))+b_{12}(t)f(x_2(t-\tau_2(t)))\\ x_2^{'}(t) &= -c_2(t)x_2(t)+a_{21}(t)f(x_1(t))+a_{22}(t)f(x_2(t))\\ &\quad +b_{21}(t)f(x_1(t-\tau_1(t)))+b_{22}(t)f(x_2(t-\tau_2(t)))\end{aligned} \tag{9}$$

where the activation function is $f_i(x)=\tanh x$. Clearly, $f_i(x)$ satisfy hypothesis (H) above , with $L_1=L_2=1$. For model (9), taking

$$\begin{aligned}c_1(t) &= e^t+2|\sin t|+3,\ c_2(t)=|\sin t|+|\cos t|+3;\\ a_{11}(t) &= e^t+|\sin t|,\ a_{12}(t)=\sin t;\\ a_{21}(t) &= \sin t,\ a_{22}(t)=-\cos t;\\ b_{11}(t) &= 1+\sin t,\ b_{12}(t)=1-\cos t;\\ b_{21}(t) &= 1-\sin t,\ \ b_{22}(t)=1+\cos t;\\ \tau_1(t) &= \tau_2(t)=\frac{1}{2}\left(|t+1|-|t-1|\right);\\ \alpha_1 &= \alpha_2=1,\end{aligned}$$

then we can easily check that

$$\begin{aligned}&k_1(t)=3; k_2(t)=2\\ &\alpha=\inf_{t\geq t_0}\{k_1(t)-k_2(t)\}=1>0\end{aligned}$$

Therefore, it follows from Theorem 1 that the system (9) is globally exponentially stable.

Remark 2. Since the delay function $\tau(t)$ in Eq.(9) is not differentiable, the results in [3] and in [22] can not be applied to this example. Furthermore, due to the unboundedness of $c_1(t)$ and $a_{11}(t)$, the results in [21] are not applicable for this example. Hence, the results here improve and extend those established earlier in [3], [22] and [21].

References

1. Arik, S.: An improved global stability result for delayed cellular neural networks. IEEE Trans.Circuits Syst.I. **49** (2002) 1211–1214
2. Arik, S.: An analysis of global asymptotic stability of delayed cellular neural networks. IEEE Trans.Neural Networks. **13** (2002) 1239–1242
3. Cao, J., Wang, J.: Global asymptotic stability of a general class of recurrent neural networks with time-varying delays. IEEE Trans.Circuits Syst.I. **50** (2003) 34–44
4. Chen, A., Cao, J., Huang, L.: An estimation of upperbound of delays for global asymptotic stability of delayed Hopfiled neural networks. IEEE Trans.Circuits Syst.I. **49** (2002) 1028–1032
5. Chua, L.O., Yang, L.: Cellular neural networks:theory and applications. IEEE Trans.Circuits Syst.I. **35** (1988) 1257–1290
6. Feng, C.H., Plamondon, R.: On the stability analysis of delayed neural networks systems. Neural Networks. **14** (2001) 1181–1188
7. Huang, H., Cao, J.: On global asymptotic stability of recurrent neural networks with time-varying delays. Appl.Math.Comput. **142** (2003) 143–154
8. Liao, X., Chen, G., Sanchez, E.N.: LMI-based approach for asymptotically stability analysis of delayed neural networks. IEEE Trans.Circuits Syst.I. **49** (2002) 1033–1039
9. Liao, X.X., Wang, J.: Algebraic criteria for global exponential stability of cellular neural networks with multiple time delays. IEEE Trans.Circuits Syst.I. **50** (2003) 268–274
10. Mohamad, S., Gopalsamy, K.: Exponential stability of continuous-time and discrete-time cellular neural networks with delays. Appl.Math.Comput. **135** (2003) 17–38
11. Roska, T., Wu, C.W., Chua, L.O.: Stability of cellular neural network with dominant nonlinear and delay-type templates. IEEE Trans.Circuits Syst.**40** (1993) 270–272
12. Zeng, Z., Wang, J., Liao, X.: Global exponential stability of a general class of recurrent neural networks with time-varying delays. IEEE Trans.Circuits Syst.I. **50** (2003) 1353–1358
13. Zhang, J.: Globally exponential stability of neural networks with variable delays. IEEE Trans.Circuits Syst.I. **50** (2003) 288–290
14. Zhang, Q., Ma, R., Xu, J.: Stability of cellular neural networks with delay. Electron. Lett. **37** (2001) 575–576
15. Qiang, Z., Ma, R., Chao, W., Jin, X.: On the global stability of delayed neural networks. IEEE Trans. Automatic Control **48** (2003) 794–797
16. Zhang, Q., Wei, X.P. Xu, J.: Global Exponential Convergence Analysis of Delayed Neural Networks with Time-Varying Delays. Phys.Lett.A **318** (2003) 537–544
17. Zhang, Q., Wei, X.P. Xu, J.: Global Asymptotic Stability of Hopfield Neural Networks with Transmission Delays. Phys.Lett.A **318** (2003) 399–405

18. Zhang, Q., Wei, X.P. Xu, J.: An Analysis on the Global Asymptotic Atability for Neural Networks with Variable Delays. Phys.Lett.A **328** (2004) 163–169
19. Zhang, Q., Wei, X.P. Xu, J.: On Global Exponential Stability of Delayed Cellular Neural Networks with Time-Varying Delays. Appl.Math.Comput. **162** (2005) 679–686
20. Zhang, Q., Wei, X.P. Xu, J.: Delay-Dependent Exponential Stability of Cellular Neural Networks with Time-Varying Delays. Chaos, Solitons & Fractals **23** (2005) 1363–1369
21. Zhou, D., Cao, J.: Globally exponential stability conditions for cellular neural networks with time-varying delays. Appl.Math.Comput. **131** (2002) 487–496
22. Jiang, H., Li, Z., Teng, Z.: Boundedness and stability for nonautonomous cellular neural networks with delay. Phys.Lett.A **306** (2003) 313–325
23. Hou, C., Qian, J.: Remarks on quantitative analysis for a family of scalar delay differential inequalities. IEEE Trans. Automatic Control **44** (1999) 334–336

Part II

Syntactical Pattern Recognition

Comparison of Two Different Prediction Schemes for the Analysis of Time Series of Graphs

Horst Bunke[1], Peter Dickinson[2], and Miro Kraetzl[2]

[1] Institut für Informatik und angewandte Mathematik,
Universität Bern, Neubrückstrasse 10, CH-3012 Bern, Switzerland
bunke@iam.unibe.ch

[2] Intelligence Surveillance Reconnaissance Division, Defence Science and Technology Organisation, Edinburgh SA 5111, Australia
{Peter.Dickinson,Miro.Kraetzl}@dsto.defence.gov.au

Abstract. This paper is concerned with time series of graphs and compares two novel schemes that are able to predict the presence or absence of nodes in a graph. Our work is motivated by applications in computer network monitoring. However, the proposed prediction methods are generic and can be used in other applications as well. Experimental results with graphs derived from real computer networks indicate that a correct prediction rate of up to 97% can be achieved.

1 Introduction

Time series, or sequence, data are encountered in many applications, such as financial engineering, audio and video databases, biological and medical research, and weather forecast. Consequently, the analysis of time series has become an important area of research [1]. Particular attention has been paid to problems such as time series segmentation [2], retrieval of sequences or partial sequences [3], indexing [4], classification of time series [5], detection of frequent subsequences [6], periodicity detection [7] and prediction [8–10].

Typically a time series is given in terms of symbols, numbers, or vectors [1]. In the current paper we go one step further and consider time series of graphs. A time series of graphs is a sequence, $s = g_1, \ldots, g_n$, where each g_i is a graph. In a recent survey it has been pointed out that graphs are a very suitable and powerful data structure for many operations needed in data mining in intelligent information processing [11]. As a matter of fact, traditional data structures, such as sequences of symbols, numbers, or vectors, can all be regarded as a special case of sequences of graphs.

The work presented in this paper is motivated by one particular application, which is computer network monitoring. In this application, graphs play an important role [12]. The basic idea is to represent a computer network by a graph, where the clients and servers are modeled by nodes and physical connections correspond to edges. If the state of the network is captured at regular points in time

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 99–106, 2005.

and represented as a graph, a sequence, or time series, of graphs is obtained that formally represents the network. Given such a sequence of graphs, abnormal network events can be detected by measuring the dissimilarity, or distance, between a pair of graphs that represent the network at two consecutive points in time. Typically an abnormal event manifests itself through a large graph distance [12].

In the current paper we address the problem of recovering incomplete network knowledge. Due to various reasons it may happen that the state of a network node or a network link can't be properly captured during network monitoring. This means that it is not known whether a certain node or edge is actually present or not in the graph sequence at a certain point in time. In this paper we compare two different schemes to recover missing information of this kind. The first procedure uses context in time, i.e. the past behaviour of a node is used to decide about its presence or absence in the present graph. By contrast, the second procedure uses within-graph context, which means that the decision about the presence or absence of the current node is based on the presence or absence of other nodes in the same graph. An information recovery procedure as described in this paper can also be used to predict, at time t, whether a certain computer in the network or a certain link will be present, i.e. active, at the next point in time, $t+1$. Such procedures are useful in computer network monitoring in situations where one or more network probes have failed. Here the presence, or absence, of certain nodes and edges is not known. In these instances, the network management system would be unable to compute an accurate measurement of network change. The techniques described in this paper can be used to determine the likely status of this missing data and hence reduce false alarms of abnormal change. Although the motivation of our work is in computer network monitoring the methods described in this paper are fairly general and can be applied in other domains as well.

The rest of this paper is organized as follows. Basic terminology and notation will be introduced in the next section. Then, in Sections 3 and 4 we will describe our two novel information recovery and prediction schemes. Experimental results will be presented in Section 5 and conclusions drawn in Section 6.

2 Basic Concepts and Notation

A labeled graph is a 4-tuple, $g = (V, E, \alpha, \beta)$, where V is the finite set of nodes, $E \subseteq V \times V$ is the set of edges, $\alpha : V \rightarrow L$ is the node labeling function, and $\beta : E \rightarrow L^{'}$ is the edge labeling function, with L and $L^{'}$ being the set of node and edge labels, respectively. In this paper we focus our attention on a special class of graphs that are characterized by unique node labels. That is, for any two nodes, $x, y \in V$, if $x \neq y$ then $\alpha(x) \neq \alpha(y)$. Properties of this class of graphs have been studied in [13]. In particular it has been shown that problems such as graph isomorphism, subgraph isomorphism, maximum common subgraph, and graph edit distance computation can be solved in time that is only quadratic in the number of nodes of the larger of the two graphs involved.

To represent graphs with unique node labels in a convenient way, we drop set V and define each node in terms of its unique label. Hence a graph with unique node labels can be represented by a 3-tuple, $g = (L, E, \beta)$ where L is the set of node labels occurring in g, $E \subseteq L \times L$ is the set of edges, and $\beta : E \rightarrow L^{'}$ is the edge labeling function [13]. The terms "node label" and "node" will be used synonymously in the remainder of this paper.

In this paper we will consider time series of graphs, i.e. graph sequences, $s = g_1, g_2, \ldots, g_N$. The notation $g_i = (L_i, E_i, \beta_i)$ will be used to represent individual graph g_i in sequence s; $i = 1, \ldots, N$. Motivated by the computer network analysis application considered in this paper, we assume the existence of a universal set of node labels, or nodes, $\mathcal{L}$, from which all node labels that occur in a sequence s are drawn. That is, $L_i \subseteq \mathcal{L}$ for $i = 1, \ldots, N$ and $\mathcal{L} = \bigcup_{i=1}^{N} L_i$.[1]

Given a time series of graphs, $s = g_1, g_2, \ldots, g_N$, and its corresponding universal set of node labels, $\mathcal{L}$, we can represent each graph, $g_i = (L_i, E_i, \beta_i)$, in this series as a 3-tuple $(\gamma_i, \delta_i, \widehat{\beta_i})$ where

- $\gamma_i : \mathcal{L} \rightarrow \{0, 1\}$ is a mapping that indicates whether node l is present in g_i or not. If l is present in g_i, then $\gamma_i\left(l\right) = 1$; otherwise $\gamma_i\left(l\right) = 0$.[2]
- $\delta_i : \mathcal{L}^{'} \times \mathcal{L}^{'} \rightarrow \{0, 1\}$ is a mapping that indicates whether edge (l_1, l_2) is present in g_i or not; here we choose $\mathcal{L}^{'} = \{l \mid \gamma_i\left(l\right) = 1\}$, i.e. $\mathcal{L}^{'}$ is the set of nodes that are actually present in g_i.
- $\widehat{\beta_i} : \mathcal{L}^{'} \times \mathcal{L}^{'} \rightarrow L^{'}$ is a mapping that is defined as follows:
  $$\widehat{\beta_i}\left(e\right) = \begin{cases} \beta_i\left(e\right), & \text{if } e \in \{(l_1, l_2) \mid \delta_i\left(l_1, l_2\right) = 1\} \\ \text{undefined}, & \text{otherwise} \end{cases}$$

The definition of $\widehat{\beta_i}\left(e\right)$ means that each edge e that is present in g_i will have label $\beta_i\left(e\right)$. The 3-tuple $(\gamma_i, \delta_i, \widehat{\beta_i})$ that is constructed from $g_i = (L_i, E_i, \beta_i)$ will be called the *characteristic representation* of g_i, and denoted by $\chi\left(g_i\right)$. Clearly, for any given graph sequence $s = g_1, g_2, \ldots, g_N$ the corresponding sequence $\chi\left(s\right) = \chi\left(g_1\right), \chi\left(g_2\right), \ldots, \chi\left(g_N\right)$ can be easily constructed and is uniquely defined. Conversely, given $\chi\left(s\right) = \chi\left(g_1\right), \chi\left(g_2\right), \ldots, \chi\left(g_N\right)$ we can uniquely reconstruct $s = g_1, g_2, \ldots, g_N$.

In the current paper we'll pay particular attention to graph sequences with missing information. There are two possible cases of interest. First it may not be known whether node l is present in graph g_i or not. In other words, in $\chi\left(g_i\right)$ it is not known whether $\gamma_i\left(l\right) = 1$ or $\gamma_i\left(l\right) = 0$. Secondly, it may not be known whether edge (l_1, l_2) is present in g_i, which is equivalent to not knowing, in $\chi\left(g_i\right)$, whether $\delta_i\left(l_1, l_2\right) = 1$ or $\delta_i\left(l_1, l_2\right) = 0$. To cope with the problem of missing information, we extend functions γ and δ in the characteristic representation, $\chi\left(g\right)$, of graph $g = (L, E, \beta)$ by including the special symbol ? in the range of

[1] In the computer network analysis application $\mathcal{L}$ will be, for example, the set of all unique IP host addresses in the network. Note that in one particular graph, g_i, usually only a subset is actually present. In general, $\mathcal{L}$ may be any finite or infinite set.

[2] One can easily verify that $\{l \mid \gamma_i\left(l\right) = 1\} = L_i$.

values of each function to indicate the case of missing information. That is, we write $\gamma(l) = ?$ if it is unknown whether node l is present in g or not.

3 Recovery of Missing Information Using Context in Time

We assume the existence of a reference set, R, of graph subsequences of length M. The reference set is defined as $R = \{s_1, \ldots, s_n\}$ where $s_j = g_{j,1}, \ldots, g_{j,M}$ for $j = 1, \ldots, n$. Each element, s_j, of the reference set is a sequence of graphs of length M. These sequences are used to represent information about the "typical behaviour" of the nodes and edges in a graph sequence of length M. This information will be used to make a decision as to $\gamma_t(l) = 0$ or $\gamma_t(l) = 1$ whenever $\gamma_t(l) = ?$ occurs.

To generate reference set R, we can utilize graph sequence $g_1, \ldots, g_{t-1}$. Each sequence in R is of length M, by definition. Let's assume that $M \leq t-1$. Then we can extract all subsequences of length M from sequence $g_1, \ldots, g_{t-1}$, and include them in reference set R. This results in

$$R = \{s_1 = g_1, \ldots, g_M \ ; \ s_2 = g_2, \ldots, g_{M+1} \ ; \ s_{t-M} = g_{t-M}, \ldots, g_{t-1}\}.$$

From each sequence, $s_i = g_i, \ldots, g_{i+M-1}$, in set R we can furthermore extract, for each node $l \in \mathcal{L}$, the sequence $\gamma_i(l), \ldots, \gamma_{i+M-1}(l)$. Assume for the moment that $\gamma_i(l), \ldots, \gamma_{i+M-1}(l) \in \{0,1\}$, which means that none of the elements $\gamma_i(l), \ldots, \gamma_{i+M-1}(l)$ is equal to ?. Then $(\gamma_i(l), \ldots, \ \gamma_{i+M-1}(l))$ is a sequence of binary numbers, 0 or 1, that indicate whether or not node l occurs in a particular graph in sequence s_i. Such a sequence of binary numbers will be called a *reference pattern.* Obviously $(\gamma_i(l), \ldots, \gamma_{i+M-1}(l)) \in \{0,1\}^M$. Because there are 2^M different binary sequences of length M, there exist at most 2^M different reference patterns for each node $l \in \mathcal{L}$. Note that a particular reference pattern, $x = (x_1, \ldots, x_M) \in \{0,1\}^M$, may have multiple occurrences in set R.

In order to make a decision as to $\gamma_t(l) = 0$ or $\gamma_t(l) = 1$, given $\gamma_t(l) = ?$, the following procedure can be adopted. First, we extract from graph sequence $s = g_1, \ldots, g_t$ the sequence $(\gamma_{t-M+1}(l), \ldots, \gamma_t(l))$ where, according to our assumption, $\gamma_t(l) = ?$. Assume furthermore that $\gamma_{t-M+1}(l), \ldots, \gamma_t(l) \in \{0,1\}$, i.e. none of the elements in sequence $(\gamma_{t-M+1}(l), \ldots, \gamma_t(l))$, except $\gamma_t(l)$, is equal to ?. Sequence $(\gamma_{t-M+1}(l), \ldots, \gamma_t(l))$ will be called the *query pattern.* Given the query pattern, we retrieve from the reference set, R, all reference patterns $x = (x_1, \ldots, x_M)$ where $x_1 = \gamma_{t-M+1}(l), x_2 = \gamma_{t-M+2}(l), \ldots, x_{M-1} = \gamma_{t-1}(l)$. Any reference pattern, x, with this property is called a *matching reference pattern.* Clearly, a reference pattern that matches the query pattern is a sequence of 0's and 1's of length M, where the first $M-1$ elements are identical to corresponding elements in the query pattern. The last element in the query pattern is equal to ?, by definition, while the last element in any matching reference pattern is either 0 or 1. Let k be the number of reference patterns that match the query pattern. Furthermore, let k_0 be the number of matching reference patterns with $x_M = 0$, and k_1 be the number of matching reference patterns with $x_M = 1$; note that $k = k_0 + k_1$. Now we can apply the following decision rule:

$$\gamma_t(l) = \begin{cases} 0 \text{ if } k_0 > k_1 \\ 1 \text{ if } k_1 > k_0 \end{cases}$$

In case $k_0 = k_1$ a random decision is in order. Intuitively, under this decision rule we consider the history of node l over a time window of length M and retrieve all cases recorded in set R that match the current history. Then a decision is made as to $\gamma_t(l) = 0$ or $\gamma_t(l) = 1$, depending on which case occurs more frequently in the reference set.

This method is based on the assumption that none of the reference patterns for node l, extracted from set R, contains the symbol ?. For a generalization where this restriction is no longer imposed neither in the reference nor in the query patterns, see [14].

4 Recovery of Missing Information Using Within-Graph Context

In this section we describe another procedure for information recovery that uses within-graph context, i.e. the presence or absence of a node is determined based on the presence or absence of other node in the same graph. The proposed procedure is based on decision trees. For all technical details of decision trees and decision tree learning see [15]. Our goal is to make a decision as to $\gamma_t(l) = 0$ or $\gamma_t(l) = 1$, given $\gamma_t(l) = ?$. Actually, this decision problem can be transformed into a classification problem as follows. The network at time t, g_t, corresponds to the unknown object to be classified. Network g_t is described by means of a feature vector, $\mathbf{x} = (x_1, \ldots, x_d)$, and the decision as to $\gamma_t(l) = 0$ or $\gamma_t(l) = 1$ can be interpreted as a two-class classification problem, where $\gamma_t(l) = 0$ corresponds to class Ω_0 and $\gamma_t(l) = 1$ corresponds to class Ω_1. As features $x_1, \ldots, x_d$ that represent the unknown object $\mathbf{x}$, i.e. graph g_t, one can use, in principle, any quantity that is extractable from graphs $g_1, \ldots, g_t$. In this paper we consider the case where these features are extracted from graph g_t exclusively. Assume that the universal set of node labels is given by $\mathcal{L} = \{l_0, l_1, \ldots, l_D\}$, and assume furthermore that it is node label l_0 for which we want to make a decision as to $\gamma_t(l_0) = 0$ or $\gamma_t(l_0) = 1$, given $\gamma_t(l_0) = ?$. Then we set $d = D$ and use the D-dimensional binary feature vector $(\gamma_t(l_1), \ldots, \gamma_t(l_D))$ to represent graph g_t. In other words, $\mathbf{x} = (\gamma_t(l_1), \ldots, \gamma_t(l_D))$. This feature vector is to be classified as either belonging to class Ω_0 or Ω_1. The former case correspond to deciding $\gamma_t(l_0) = 0$, and the latter to $\gamma_t(l_0) = 1$. Intuitively, using $(\gamma_t(l_1), \ldots, \gamma_t(l_D))$ as a feature vector for the classification of g_t means we make a decision as to the presence or absence of l_0 in g_t depending on the presence or absence of all other nodes from $\mathcal{L}$ in g_t. The classification procedure actually implemented is a decision tree [15]. For further details see [14].

5 Experimental Results

The methods described in Sections 3 and 4 of this paper have been implemented and experimentally evaluated on real network data. For the experiments four

Table 1. Characterisation of the graph sequences used in the experiments.

	S_1	S_2	S_3	S_4
Number of graphs in sequence	102	292	202	99
Size of smallest graph in sequence	38	85	15	572
Size of largest graph in sequence	94	154	329	10704
Average size of graphs in sequence	69.7	118.5	103.9	5657.8

time series of graphs, S_1, S_2, S_3 and S_4, acquired from existing computer networks have been used. Characteristics of these graph sequences are shown in Table 1, where the size of a graph is defined as the number of its nodes. All four series represent logical communications on the network. Series S_1, S_2 and S_4 were derived from data collected from a large enterprise data network, while S_3 was collected from a wireless LAN used by delegates during the World Congress for Information Technology (WCIT2002). The nodes in each graph of S_1 and S_2 represent business domains in the network, while in S_3 and S_4 they represent individual IP addresses. Note that all graph sequences are complete, i.e. there are no missing nodes and edges in these sequences.

To test the ability of the method described in Section 3 it was assumed, for each graph in a time series, that $\gamma(l)$ is unknown for each node. Then the prediction scheme was applied and the percentage of correctly predicted nodes in each graph of the sequence was determined. In some preliminary experiments it was found out that the optimal size of the time window is $M = 5$. Hence this value was used. In Fig. 1 the percentage of correctly predicted nodes for each graph of sequence S_1 is shown.

To test the method described in Section 4, each time series is divided into two disjoined sets of graphs. The first set, G_1, consists of all graphs g_i with index i being an odd number (i.e. graphs $g_1, g_3, g_5, \ldots$), while the other set, G_2, includes all graphs with an even index i (i.e. graphs $g_2, g_4, \ldots$). First, set G_1 is used as a training set for decision tree induction and G_2 serves as a test set. Then G_1 and G_2 change their role, i.e. G_1 becomes the test and G_2 the training set. For each graph, g, in the test set we count the number of nodes that have been correctly predicted and divide this number by the total number of nodes in g. The correct prediction rate obtained with this method is also shown in Fig. 1. We observe that for both methods the correct prediction rate is typically in the range of $85\% - 95\%$. This is a remarkably high value taking into consideration that for a two-class classification problem, such as the one considered here, random guessing would give us an expected performance of only 50%.

Because of space limitations, we only show the results for sequence S_1. However the results for the other three time series are very similar. A summary of all our experimental results is provided in Table 2, where the correct prediction rate is averaged over all graphs in a sequence.

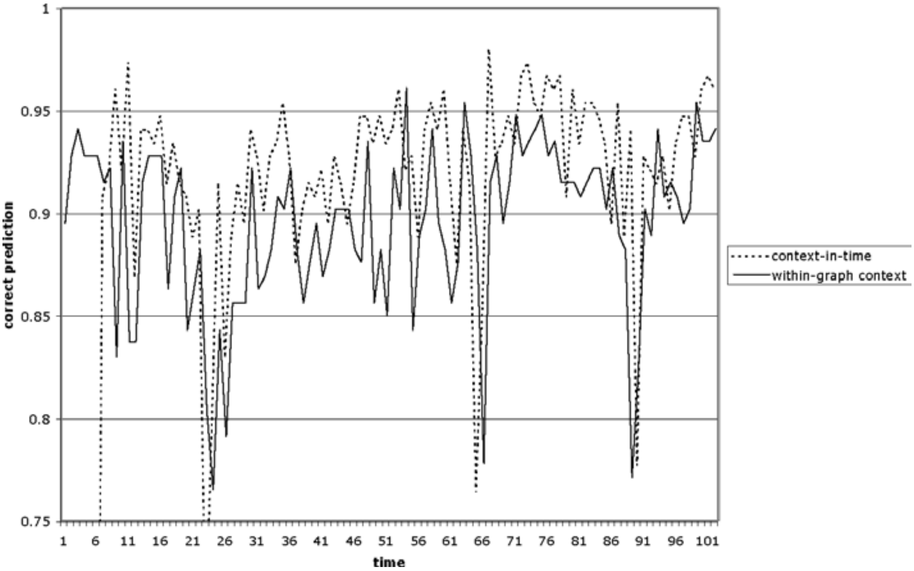


Fig. 1. Correct prediction rates for sequence S_1 obtained with context-in-time and within-graph context method.

Table 2. Summary of experimental results.

	S_1	S_2	S_3	S_4
Context in Time (Sec. 3)	92.1	97.2	90.5	95.1
Context within Graph (Sec. 4)	89.5	93.4	96.9	89.4

6 Conclusions

The problem of incomplete knowledge recovery and prediction of the behaviour of nodes in time series of graphs is studied in this paper. Formally, this task is formulated as a classification problem where nodes and edges with an unknown status are to be assigned to one of the classes 'present in' or 'absent from' the actual graph. Two different schemes are proposed in order to solve this classification problem. One of these schemes uses context in time, while the other is based on context within the actual graph. Both procedures achieve impressive prediction rates up to about 97% on sequences of graphs derived from real computer network data. The motivation of this work derives from the field of computer network monitoring. However the proposed framework for graph sequence analysis is fairly general and can be applied in other domains as well.

Acknowledgement

The authors are thankful to Ch. Irniger, J. Marbach, F. Thalmann and T. Varga for valuable contributions to this paper.

References

1. Last, M., Kandel, A., Bunke, H. (eds.): *Data Mining in Time Series Databases*, World Scientific, 2004
2. Keogh, E. et al: Segmenting time series: A survey and novel approach, in [1], 1-21
3. Kahveci, T., Singh, K.: Optimizing similarity search for arbitrary length time series queries, IEEE Trans. KDE, Vol 16, No 2, 2004, 418-433
4. Vlachos, M. et al.: Indexing time-series under conditions of noise, in [1], 67-100
5. Zeira, G. et al: Change detection in classification models induced from time series data, in [1], 101-125
6. Tanaka, H., Uehara, K.: Discover motifs in multi-dimensionaltime-series using the principle component analysis and the MDL principle, in Perner, P., Rosenfeld, A. (eds.): Machine Learning and Data Mining in Pattern Recognition, *Proc. 3rd Int. Conference*, Springer LNAI 2734, 2003, 252-265
7. Yang, J., Wang, W., Yu, P.S.: Mining asynchronous periodic patterns in time series data, IEEE Trans. KDE, Vol. 15, No 3, 2003, 613-628
8. Schmidt, R., Gierl, L.: Temporal abstractions and case-based reasoning for medical course data: two prognostic applications, in Perner, P. (ed.): *Machine Learning in Pattern Recognition Proc. 2nd Int. Workshop*, Springer, LNAI 2123, 2001, 23-34
9. Fung, G.P.C.F., Yu, D.X., Lam, W.: News sensitive stock trend prediction, in Chen, M.-S., Yu, P.S., Liu, B. (eds.): *Advances in Knowledge Discovery and Data Mining, Proc. 6th Pacific-Asia Conference, PAKDD*, Springer, LNAI 2336, 2002, 481-493
10. Povinelli, R.F., Feng, X.: A new temporal pattern identification method for characterization and prediction of complex time series events, IEEE Trans. KDE, Vol. 15, No 2, 2003, 339-352
11. Bunke, H.: Graph-based tools for data mining and machine learning, in P. Perner, A. Rosenfeld (eds.): Machine Learning and Data Mining in Pattern Recognition, *Proc. 3rd Int. Conference*, Springer LNAI 2734, 2003, 7-19
12. Bunke, H., Kraetzl, M., Shoubridge, P., Wallis, W.: Detection of abnormal change in time series of graphs, *Journal of Interconnection Networks*, Vol. 3, Nos 1,2, 2002, 85-101
13. Dickinson, P., Bunke, H., Dadej, A., Kraetzl, M.: Matching graphs with unique node labels, accepted for publication in *Pattern Analysis and Applications*
14. Bunke, H., Dickinson, P., Kraetzl, M.: Analysis of Graph Sequences and Applications to Computer Network Monitoring, *Technical Report*, DSTO, Edinburgh, Australia, 2003
15. Quinland, R.: C4.5: Programs for Machine Learning, Morgen Kaufmann Publ., 1993

Grouping of Non-connected Structures by an Irregular Graph Pyramid*

Walter G. Kropatsch and Yll Haxhimusa

Pattern Recognition and Image Processing Group 183/2,
Institute for Computer Aided Automation,
Vienna University of Technology, Austria
{krw,yll}@prip.tuwien.ac.at

Abstract. Motivated by claims to 'bridge the representational gap between image and model features' and by the growing importance of topological properties we discuss several extensions to dual graph pyramids: structural simplification should preserve important topological properties and content abstraction could be guided by an external knowledge base. We review multilevel graph hierarchies under the special aspect of their potential for abstraction and grouping.

1 Introduction

Regions as aggregations of primitive pixels play an extremely important role in nearly every image analysis task. Regional (internal) properties (color, texture, shape, ...) help to identify them and their external relations (adjacency, inclusion, similarity of properties,...) are used to build groups of regions having a particular meaning in a more abstract context. A question is raised in [11] referring to several research issues: "How do we bridge the representational gap between image features and coarse model features?" They identify the 1-to-1 correspondence between: *salient image features* (pixels, edges,...) and *salient model features* (generalized cylinders, invariant models,...) as *limiting assumption* that makes generic object recognition impossible. It is suggested to *bridge* and not to *eliminate the representational gap*, and to focus efforts on: region segmentation, *perceptual grouping* and *image abstraction*.

Connected components form the bases for most segmentations. The region adjacency graph (RAG) describes the relations of connected regions. However not all regions of the RAG have the same importance like a dotted line on white background. In such cases the more important regions are offten close to each other but not adjacent and adjacency prevents further grouping. We overcome this problem by letting more important regions (foreground) grow into the non important regions (background) until the close regions become adjacent and can be grouped. We address some of these issues in the context of gradually generalizing our discrete image data across levels where geometry dominates up to levels of the hierarchy where topological properties become important.

* Supported by the Austrian Science Foundation under grant FSP-S9103-N04.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 107–114, 2005.

We review the formal definition of abstraction (Sec. 2) and the concept of dual graphs (Sec. 3) including a ‘natural’ example of vision based on an irregular sampling. Image pyramids of dual graphs are the main focus of Sec. 4. Abstraction in multilevel structures can be done either by modifying the contents of a representational cell or by ‘simplifying’ the structural arrangement of the cells while major topological properties are preserved (Sec. 5).

2 Visual Abstraction

By definition abstraction extracts essential features and properties while it neglects unnecessary details. Two types of unnecessary details can be distinguished: *redundancies* and *data of minor importance*. Details may not be necessary in different contexts and under different objectives which reflect in different types of abstraction. In general we distinguish: *isolating abstraction*: important aspects of one or more objects are extracted from their original context; *generalizing abstraction*: typical properties of a collection of objects are emphasized and summarized. *idealizing abstraction*: data are classified into a (finite) set of ideal models, with parameters approximating the data and with (symbolic) names/notions determining their semantic meaning. These three types of abstraction have strong associations with well known tasks in cognitive vision: recognition and object detection tries to *isolate* the object from the background; perceptual grouping needs a high degree of *generalization*; and categorization *assigns* data to *ideal classes* disregarding noise and measurement inaccuracies. In all cases abstraction drops certain data items which are considered less relevant. Hence the *importance* of the data needs to be computed to decide which items to drop during abstraction. The importance or the relevance of an entity of a (discrete) description must be evaluated with respect to the purpose or the goal of processing.

Multiresolution hierarchies, image pyramids or trees in general posses the potential for abstraction. We consider the structure of the representation and the content stored in the representational units separately. In our generalization we allow the resolution cell to take other simply connected shapes and to describe the content by a more complex ‘language’. The first generalization is a consequent continuation of the observations in [2] to overcome the limited representational capabilities of rigid regular pyramids. Since irregular structures reduce the importance of explicitly representing geometry, topological aspects become relevant.

3 Discrete Representation – Dual Graphs

A digital image is a finite subset of ‘pixels’ of the discrete grid $\mathbb{Z}^2$. The discretization process maps any object of the continuous image into a discrete version if it is sufficiently large to be captured by the sensors at the sampling points. Resolution relates the unit distance of the sampling grid with a distance in reality. The properties of the continuous object, i.e. color, texture, shape, as well as its relations to other (nearby) objects are mapped into the discrete space,

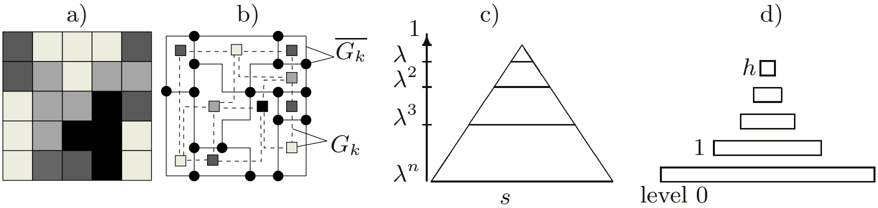


Fig. 1. a) Partition of pixel set into cells. b) Representation of the cells and their neighborhood relations ($G_k, \overline{G_k}$). c) Pyramid concept, and d) discrete levels.

too. The most primitive discrete representation assigns to each sampling point a measurement, be it a gray, color or binary value. In order to express the connectivity or other geometric or topological properties, the discrete representation must be enhanced by a neighborhood relation. In the regular square grid 4- or 8-neighborhood have the well known problems in conjunction with Jordan's curve theorem. The neighborhood of sampling points is represented by a graph. Although this data structure consumes more memory space it has several advantages, among which we find the following: *the sampling points need not be arranged in a regular grid*; *the edges can receive additional attributes too*; and *the edges may be determined either automatically or depending on the data.*

The problem arising with irregular grids is that there is no implicit neighbor definition. Usually Voronoi neighbors determine the neighborhood graph. The neighborhood in irregular grids needs to be represented explicitly. This creates a new representational entity: the binary relation of an edge in the neighborhood graph similar to the concept of relations between observational entities in [5]. Together with the fact that a $2D$ image is embedded in the continuous image plane, the line segments connecting the end points of edges partition the image plane into connected faces which are part of the *dual graph* (Fig. 1a,b).

4 Pyramids

In this section we summarize the concepts developed for building and using multiresolution pyramids [10, 15] and put the existing approaches into a general framework. The focus of the presentation is a representational framework, its components and the processes that transfer data within the framework. A pyramid [15] (Fig. 1c,d) describes the contents of an image at multiple levels of resolution. The base level is a high resolution input image. Successive levels reduce the size of the data by a constant *reduction factor* $\lambda > 1.0$ while local *reduction windows* relate one cell at the reduced level with a set of cells in the level directly below. Thus local independent (and parallel) processes propagate information up and down in the pyramid. The contents of a lower resolution cell is computed by means of a *reduction function*, the input of which are the descriptions of the cells in the reduction window.

The number of levels n is limited by λ: $n \leq \log(image\_size)/\log(\lambda)$. The main computational advantage of *image pyramids* is due to this *logarithmic com-*

plexity. We intend to extend the expressive power of these efficient structures by several generalizations. In order to interpret a derived description at a higher level, this description should be related to the original input data in the base of the pyramid. The *receptive field* (RF) of a given pyramidal cell c_i, $RF(c_i)$, collects all cells (pixels) in the base level of which c_i is the ancestor.

Content Models and Reduction Functions

In connected component labeling each cell contains a label identifying the membership of the cell to the class of all those cells having the same label. In this case the contents of the cells merged during the reduction process can be propagated by simple inheritance: the fused cell 'inherits' its label from its children. In classical gray level pyramids the contents of a cell is a gray value which is summarized by the mean or a weighted mean of the values in the reduction window. Such reduction functions have been used in Gaussian pyramids. Laplacian pyramids [4] and wavelet pyramids [16] identify the loss of information that occurs in the reduced level and store the missing information in the hierarchical structure where it can be retrieved when the original is reconstructed. These approaches use *one single globally defined model* [8] which must be flexible to adapt its parameters to approximate the data.

In our generalization we would like to go one step further and allow *different models* to be used in different resolution cells as there are usually different objects *at different locations* of an image. The models could be identified by a name or a symbol (e.g black, white, isolated etc.) and may be interrelated by semantic constraints (e.g adjacency etc.), Fig. 4. Simple experiments have been done with images of line drawings. This research used the experiences gained with a system for perceptional curve tracing based on regular $2\times2/2$ curve pyramid [12] and the chain pyramid [17] in the more flexible framework of graph pyramids. The model describes symbolically the way in which a curve intersects the discrete segments of the boundary of a cell and the reduction function consists in the transitive closure of the symbols collected in the reduction window. The concept works well in areas where the density of curves is low, although the rigidity of the regular pyramid causes ambiguities to arise when more curves appear within the same receptive field. This limitation can be overcome with irregular pyramids [15] in which we could limit the receptive field of a cell to a single curve.

The content abstraction in this representation has following features:

- models are identified by names[1], no parameters were used;
- adjacent models have to be consistent ('good continuation');
- only one consistent curve is covered in one receptive field;
- this selection process is governed by a few contraction rules (Fig. 4).

The knowledge about the models and in what configurations they are allowed to occur needs to be stored in a knowledge base [14]. In order to determine which are the best possible abstractions, the local configurations at a given level of the pyramid must be compared with the possibilities of reduction given in the

[1] Discrete names: empty cell, line end, crosses edge, junction etc.

Algorithm 1 – Graph Pyramid.

```
Input: Attributed graph $G$.
while { further abstraction is possible } do
    determine contraction kernels (CKs),
    perform dual graph contraction and simplification of dual graph,
    apply reduction functions to compute content of new reduced level,
end while
Output: Irregular graph pyramid.
```

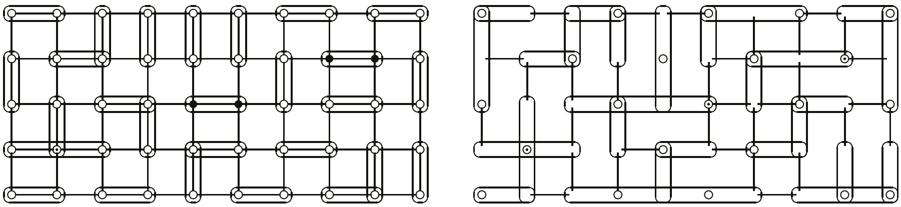
knowledge base. This would typically involve matching the local configuration with the right-hand sides of rules stored in the knowledge base. Such a match may not always be perfect, one may allow a number of outliers. The match results in a goodness of match, which can be determined for all local configurations. The selection can then choose the locally best candidates as contraction kernels (CKs) and reduce the contents according to the generic models which matched the local configuration. The goodness of match may also depend on a global objective function to allow the overall purpose, task or intention to influence the selection process.

5 Irregular Graph Pyramids

A graph pyramid is a pyramid where each level is a graph $G(V, E)$ consisting of vertices V and of edges E relating pairs of vertices. In the base level, pixels are the vertices, and two vertices are related by an edge if the two corresponding pixels are neighbors. This graph is called the neighborhood graph. The content of the graph is stored in attributes attached to both vertices and edges. In order to correctly represent the embedding of the graph in the image plane we additionally store the dual graph $\overline{G}(\overline{V}, \overline{E})$ at each level. Let us denote the original graph as the primal graph. In general a graph pyramid can be generated bottom-up [15] (see Alg. 1).

5.1 1st Iteration: Group Connected Components

The 2^{nd} step determines what information in the current top level is important and what can be dropped. A CK is a (small) sub-tree, the root of which is chosen to survive. Fig. 2a shows the window (G_0) and the selected CK $N_{0,1}$ each surrounded by an oval. The codes of the vertices are given in Fig. 4. Selection criteria (code adjacency of Fig. 4 is 'yes') in this case contract only edges inside connected components except for isolated black vertices (blobs) which are allowed to merge with their background, so that support of grouping is distributed over a large receptive field bridging areas of background [6]. All the edges of the contraction trees are dually contracted [15]. Dual contraction of an edge e (formally denoted by $G/\{e\}$) consists of contracting e and removing the corresponding dual edge $\overline{e}$ from the dual graph (formally denoted by $\overline{G} \setminus \{\overline{e}\}$).



a) Neighborhood graph G_0 and CK N_{01} b) G_1 and CK N_{12};

Fig. 2. Broken line.

This preserves duality and the dual graph need not be constructed from the contracted primal graph G' at the next level.

Since the contraction of an edge may yield multi-edges and self-loops there is a simplification step which removes all redundant multi-edges and self-loops (redundant edges). Note that not all such edges can be removed without destroying the topology of the graph since its removal would corrupt the connectivity! This can be decided locally by the dual graph since *faces of degree two* (having the double-edge as boundary) and *faces of degree one* (boundary = self-loop) cannot contain any further elements in its interior, since the original graph is connected. Since removal and contraction are dual operations, the removal of a self-loop or of one of the double edges can be done by contracting the corresponding dual edges in the dual graph. The dual contraction of our example remains a graph G_1 without redundant edges (Fig. 2b).

5.2 New Category: Isolated Blob

Step 3 generates a reduced pair of dual graphs. The content is derived in step 4 from the level below. In our example, reduction is very simple: the surviving vertex inherits the color of its son. A new category ‘isolated blob’ is introduced if a black vertex is completely surrounded by white vertices. This new label allows the RF to grow into its background and, eventually, close the gap to another isolated blob. In the only case where the CK contains two different labels, the isolated vertex is always chosen as surviving vertex.

The result of the second dual contraction is shown in Fig. 3. The selection rules and the reduction function are the same as in the first iteration. The isolated blob adjacency graph (IBAG) shows that the gaps between the isolated blobs of the original sampling have been closed and the three surviving isolated blobs are connected after two iterations. A top-down verification step checks the reliability of closing the gap. There are lots of useful properties of the resulting graph pyramids. If the plane graph is transformed into a combinatorial map the transcribed operations form the combinatorial pyramid [3]. This framework allowed to link dual graph pyramids with topological maps which extend the scope to $3D$.

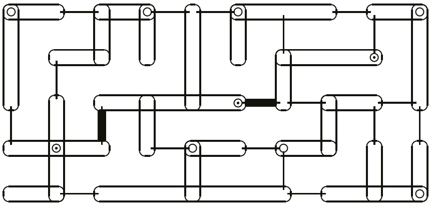
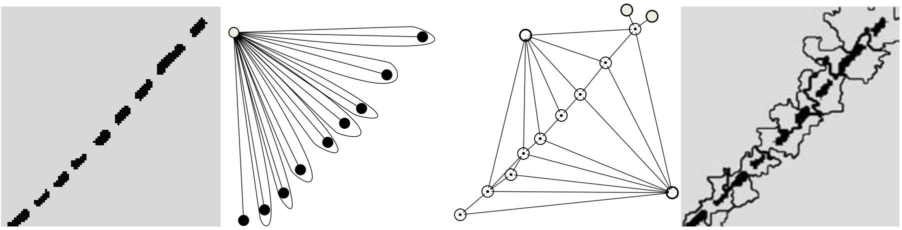


Fig. 3. The two gaps in graph G_2.

may contract with		○	●	⊙
empty background	○	yes	no	yes
black component	●	no	yes	no
isolated blob	⊙	yes	no	yes (*gap*)

Fig. 4. Contraction rules to close gaps.



a) Broken line b) CCA c) IBAG d) RF

Fig. 5. Closing the gaps of a broken line.

6 Experimental Result

Fig. 5 shows an example of closing the gaps of a broken line. Connected components analysis (CCA) alone creates self loops. Growing isolated blobs into its background produces vertices of isolated blobs connected by edges corresponding to the gaps. Fig. 5d shows the corresponding *RF* of the isolated blobs, which represent edgel hypotheses and the neighborhood of isolated vertices a line hypothesis. These hypotheses can be verified for confidence using the hierarchy of the pyramid. It seems that there are much less concepts working on discrete irregular grids than on their regular counterparts. How to group connected structures into an extended RAG has been show before [9]. The many islands of highly split structures remain isolated in these approaches. We show how to group isolated blobs or substructures into IBAG if the blobs have a 'common' background.

7 Conclusion

We motivated our discussion by the claim to *'bridge the representational gap'* [11] and to *'focus on image abstraction'*. We first discussed the basic concepts, visual abstraction and dual graphs in more detail. We then recalled a pyramidal approach having the potential to cope also with irregular grids. These pyramids have some useful properties: i) they show the need to use multi-edges and self-loops to preserve the topology; ii) they allow the combination of primitive operations at one level (i.e. collected by the CK) and across several levels of the pyramid (i.e. equivalent contraction kernels [13]); iii) repeated contraction converges to specific properties which are preserved during contraction; iv) termination criteria allow abstraction to be stopped before a certain property is lost. The new category of an isolated blob allowed to group non adjacent regions based on proximity.

References

1. Y. Bertrand, C. Fiorio, and Y. Pennaneach. Border Map: A Topological Representation for nD Image Analysis. *DGCI*, LNCS 1568:242–257, 1999.
2. M. Bister, J. Cornelis, and A. Rosenfeld. A Critical View of Pyramid Segmentation Algorithms. *Pattern Recognition Letters*, 11(9):605–617, 1990.
3. L. Brun and W. G. Kropatsch. Introduction to Combinatorial Pyramids. *Dig. and Im. Geom.*, 108–128, 2001.
4. P. J. Burt and E. H. Adelson. The Laplacian Pyramid as a Compact Image Code. *IEEE Trans. on Commun.*, 31(4):532–540, 1983.
5. J. L. Crowley, J. Coutaz, G. Rey, and P. Reignier. Perceptual Components for Context Aware Computing. In *Int. Conf. on Ubiq. Comp.*, 117–143, 2002.
6. S. C. Dakin and P. Bex. Local and Global Visual Grouping: Tuning for Spatial Frequency and Contrast. *J. of Vis.*, 99–111, 2001.
7. G. Damiand. *Définition et Étude d'un Modèle Topologique Minimal de Représentation D'images 2d et 3d*. PhD thesis, LIRMM, Uni. de Montpellier, 2001.
8. R. L. Hartley. *Multi–Scale Models in Image Analysis*. PhD thesis, Uni. of Maryland, CSC, 1984.
9. Y. Haxhimusa and W. G. Kropatsch. Hierarchical Image Partitioning with Dual Graph Contraction. *DAGM Symp.*, LNCS 2781:338–345, 2003.
10. J.-M. Jolion and A. Rosenfeld. *A Pyramid Framework for Early Vision*. Kluwer, 1994.
11. Y. Keselman and S. Dickinson. Generic model abstraction from examples. In *CVPR.*:856–863, Hawaii, 2001.
12. W. G. Kropatsch. Preserving Contours in Dual Pyramids. *ICPR*, 563–565, 1988.
13. W. G. Kropatsch. From Equivalent Weighting Functions to Equivalent Contraction Kernels. *DIP and CG2:*, SPIE 3346:310–320, 1998.
14. W. G. Kropatsch. Abstract Pyramid on Discrete Represtations. *DGCI* LNCS 2301, 1–21, 2002.
15. W. G. Kropatsch, A. Leonardis, and H. Bischof. Hierarchical, Adaptive and Robust Methods for Image Understanding. *Sur. on Math. for Ind.*, 9:1–47, 1999.
16. S. G. Mallat. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Trans. PAMI*, 11(7):674–693, 1989.
17. P. Meer, C. A. Sher, and A. Rosenfeld. The Chain Pyramid: Hierarchical Contour Processing. *IEEE Trans. PAMI*, 12(4):363–376, 1990.

An Adjacency Grammar to Recognize Symbols and Gestures in a Digital Pen Framework[*]

Joan Mas, Gemma Sánchez, and Josep Lladós

Computer Vision Center, Dept. Informàtica
Universitat Autònoma de Barcelona,
08193 Bellaterra (Barcelona), Spain
{jmas,gemma,josep}@cvc.uab.es
http://www.cvc.uab.es

Abstract. The recent advances in sketch-based applications and digital-pen protocols make visual languages useful tools for Human Computer Interaction. Graphical symbols are the core elements of a sketch and, hence a visual language. Thus, symbol recognition approaches are the basis for visual language parsing. In this paper we propose an adjacency grammar to represent graphical symbols in a sketchy framework. Adjacency grammars represent the visual syntax in terms of adjacency relations between primitives. Graphical symbols may be either diagram components or gestures. An on-line parsing method is also proposed. The performance of the recognition is evaluated using a benchmarking database of 5000 on-line symbols. Finally, an application framework for sketching architectural floor plans is described.

1 Introduction

The field of Human Computer Interaction (HCI) has new emerging interests concerned about incorporating tools from affine disciplines towards the modelization of innovative multimodal interfaces. One of these disciplines are sketching interfaces that combine pattern recognition and document analysis techniques. A sketch is a line drawing image consisting of a set of hand drawn strokes drawn by a person using an input framework. Thus, by *sketching* or *calligraphic interface* we designate those applications consisting in the use of digital-pen inputs for creation or edition of handwritten text, diagrams or drawings. A *digital pen* is like a simple ballpoint pen but uses an electronic head instead of ink. We refer to *digital ink* the chain of points obtained by a trajectory of a pen movement during touching a dynamic input device. Devices as PDAs or TabletPCs incorporate such kind of digital-pen input protocols. Interesting applications of digital-pen protocols are freehand drawing for early design stages in engineering, biometrics (signature verification), query by sketch in image database retrieval, or augmenting and editing documents. Sketching with a pen in HCI is a mode of natural, perceptual, and direct interaction in which the user has also instant feedback. Informally speaking, a sketch can be seen as a user-to-computer communication unit formulated as a valid

[*] This work has been partially supported by the spanish project CICYT TIC2003-09291 and catalan project CeRTAP PVPC.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 115–122, 2005.

sentence of a visual language. A visual language for a diagrammatic notation allows to combine elements of an alphabet of graphical symbols, i.e. bidimensional patterns, by means syntactic rules describing valid structural relations among them. According to that, grammars and parsing are very suitable tools to describe and recognize this type of patterns. Syntactic pattern recognition methods use formal grammars to describe bidimensional patterns. To do so, in the literature we can find two major approaches, namely the use of string grammars as PDL or Plex grammars [1], or grammars that are inherently bidimensional as web [2] or graph grammars [3–5].

In this work we propose a syntactic graphical symbol recognition approach in a sketchy framework. A sketchy framework can be classified depending on different criteria: regarding to input modes, it can be off-line or on-line; the information that is conveying can be text or graphics; and finally the sketched symbols can be used as freehand drawings or gestures. In this paper we focus on the category of on-line graphical sketches, either used as gestures or symbols in freehand drawings. A sketch recognition system consists of three major stages: primitive extraction, compound object recognition, and sketch understanding. In our work, primitive extraction consists in the approximation of on-line strokes by primitives as straight segments and circular arcs; compound object recognition is the syntactic stage in which symbols belonging to predefined classes are recognized in terms of an adjacency grammar; and finally sketch understanding applies semantic rules to the symbol instances recognized in the sketch. Two symbol categories has been defined one designing elements of a freehand drawing, and the other a set of gestures *delete, rotate, etc.*, used to edit the drawing elements. Since in both cases the problem consists in recognizing hand made symbols, the syntactic approach proposed in this work is common for both categories.

This paper is organized as follows: in section 2 we first introduce the adjacency grammar and afterwards the parsing process is described. Section 3 describes an experimental framework in which the proposed approach has been incorporated. Finally in section 4 we present the conclusions.

2 Recognition Process

Sketch understanding involves the recognition of graphical symbols that have a meaning in the context where they appear. We can distinguish two major symbol categories: freehand drawings or gestures. One symbol depending on whether it is drawn in one of such categories can have different meaning. Thus, in the former symbols are elements of a diagram vocabulary, and in the latter symbols have associated actions to modify the diagram. However, the recognition can be formulated in terms of a common method. Different kinds of grammars have been used in pattern recognition. In graphics recognition a grammar allows the definition of a symbol by means of a set of primitives and relations among them. In an on-line framework different users do not draw the same symbol or gesture in the same way. Therefore the recognition method should be unconstrained to the order of strokes. A particular class of grammars are *Adjacency Grammars*[6]. Since its order free nature, they are suitable to model the sketching behaviour, whilst classical grammars seem more unnatural due to the sequential organization of their production elements.

2.1 Adjacency Grammars

Adjacency grammars have been used in many disciplines to define symbols. According to the notation of [6] an adjacency grammar is formally defined as a 5-tuple $G = (V_t, V_n, S, P, C)$ where:

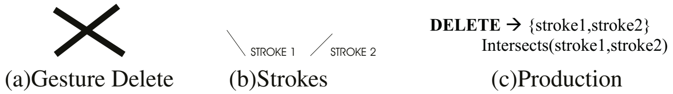
- V_t denotes the alphabet of terminal symbols.
- V_n denotes the alphabet of non-terminal symbols.
- $S \in V_n$ is the start symbol of the grammar.
- C is the set of constraints applied to the elements of the grammar.
- P are the productions of the grammar defined as:

$$\alpha \rightarrow \{\beta_1, \ldots, \beta_n\} \,\textbf{if}\, \Gamma(\beta_1, \ldots, \beta_n)$$

where: $\alpha \in V_n$ and all the $\beta_j \in \{V_t \cup V_n\}$, constitute a possibly empty multiset of terminal and non-terminal symbols. Γ is an adjacency constraint defined over the attributes of the β_j.

The symbols β_j can appear in any order, for example, in the following production: $\alpha \rightarrow \{\mu, \nu, \sigma\} \in P$ we consider all 6 possible permutations of μ, ν, σ as equally valid substitutions for α.

An example of a symbol can be seen in Fig. 1(a), the strokes forming the symbol can be seen in Fig. 1(b) and Fig. 1(c) shows the production that defines the rules to combine the strokes to form the symbol. Notice that this gesture has two strokes of type *segment* and the adjacency constraint between the strokes is of type *Intersects*.



(a)Gesture Delete (b)Strokes (c)Production

Fig. 1. Example of Gesture.

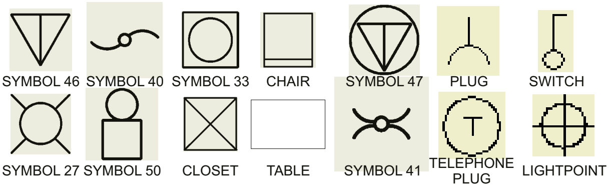
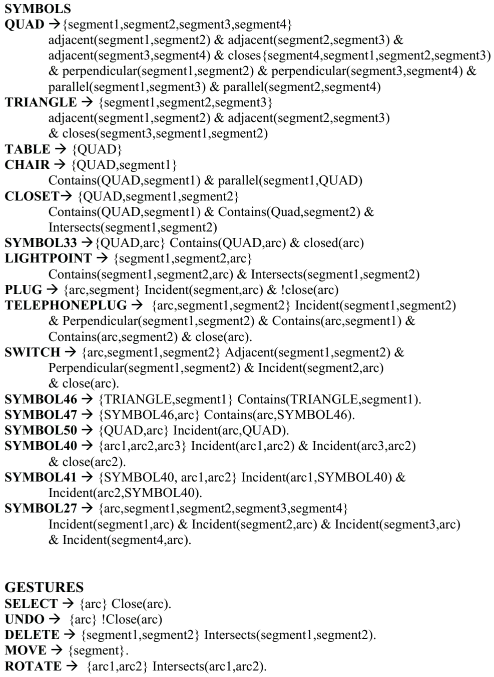


Fig. 2. Graphical symbols in an architectural domain.

In our grammar we can distinguish two levels: **lexical**, referring to the primitives extracted from the strokes forming the diagram, and **syntactic** that refers to the symbols and gestures. Let us now further describe these two levels.



```
SYMBOLS
QUAD →{segment1,segment2,segment3,segment4}
        adjacent(segment1,segment2) & adjacent(segment2,segment3) &
        adjacent(segment3,segment4) & closes{segment4,segment1,segment2,segment3)
        & perpendicular(segment1,segment2) & perpendicular(segment3,segment4) &
        parallel(segment1,segment3) & parallel(segment2,segment4)
TRIANGLE → {segment1,segment2,segment3}
        adjacent(segment1,segment2) & adjacent(segment2,segment3)
        & closes(segment3,segment1,segment2)
TABLE → {QUAD}
CHAIR → {QUAD,segment1}
        Contains(QUAD,segment1) & parallel(segment1,QUAD)
CLOSET→ {QUAD,segment1,segment2}
        Contains(QUAD,segment1) & Contains(Quad,segment2) &
        Intersects(segment1,segment2)
SYMBOL33 →{QUAD,arc} Contains(QUAD,arc) & closed(arc)
LIGHTPOINT → {segment1,segment2,arc}
        Contains(segment1,segment2,arc) & Intersects(segment1,segment2)
PLUG → {arc,segment} Incident(segment,arc) & !close(arc)
TELEPHONEPLUG →  {arc,segment1,segment2} Incident(segment1,segment2)
        & Perpendicular(segment1,segment2) & Contains(arc,segment1) &
        Contains(arc,segment2) & close(arc).
SWITCH → {arc,segment1,segment2} Adjacent(segment1,segment2) &
        Perpendicular(segment1,segment2) & Incident(segment2,arc)
        & close(arc).
SYMBOL46 → {TRIANGLE,segment1} Contains(TRIANGLE,segment1).
SYMBOL47 → {SYMBOL46,arc} Contains(arc,SYMBOL46).
SYMBOL50 → {QUAD,arc} Incident(arc,QUAD).
SYMBOL40 → {arc1,arc2,arc3} Incident(arc1,arc2) & Incident(arc3,arc2)
        & close(arc2).
SYMBOL41 → {SYMBOL40, arc1,arc2} Incident(arc1,SYMBOL40) &
        Incident(arc2,SYMBOL40).
SYMBOL27 → {arc,segment1,segment2,segment3,segment4}
        Incident(segment1,arc) & Incident(segment2,arc) & Incident(segment3,arc)
        & Incident(segment4,arc).


GESTURES
SELECT → {arc} Close(arc).
UNDO →  {arc} !Close(arc)
DELETE → {segment1,segment2} Intersects(segment1,segment2).
MOVE → {segment}.
ROTATE →  {arc1,arc2} Intersects(arc1,arc2).
```

Fig. 3. The adjacency grammar for gestures and diagram symbols in an architecture framework.

2.2 Lexical Level

This level extracts the primitive elements compounding the sketch. Primitives constitute the terminal alphabet of the grammar and they encode the strokes introduced to the system with a digital pen. It is important to distinguish between *stroke* and *primitive*. A stroke is a trajectory of a pen movement during touching a dynamic input device. A stroke is the minimal unit of user input, represented by a chain of points. A primitive refers to a simple shape that encodes either a substroke or a multistroke. Therefore, a primitive is the minimal semantic unit, i.e. a lexical token. Different alphabets of primitives can be used to encode strokes. Useful examples are polygonal approximations, dominant curvature points, or basic geometric shapes (squares, triangles, circles).

In this work, strokes are approximated by two types of primitives, namely *segment* and *arc*, i.e. $V_t = \{segment, arc\}$[1]. Since a stroke is a sequence of points, primitive extraction is formulated in terms of a vectorization process. A number of performant vectorization methods can be found in the literature, for a recent review see [7]. We will refer to this process as *on-line vectorization* because dynamic information is also used. A classical vectorization approach adjusts analytical parameters of segments and arcs in terms of a minimum-error procedure regarding to static information of image pixels.

[1] For the sake of clarity we refer as *segment1*…*segmentn* the different instances of the stroke *segment*.

Dynamic information like pressure or speed changes in the pen movement is useful to detect corner points to divide the stroke. Given a stroke, our on-line vectorization method consists of two steps. First, corner points are detected using a combination of static and dynamic information, following the idea proposed in [8]. Thus, we look for points along the stroke having maximum curvature and minimum speed. The second step, approximates each substroke between consecutive corner points by a straight segment or a circular arc in terms of a best fit criterion between the analytical primitive and the sequence of stroke points.

2.3 Syntactic Level

The syntactic level recognizes graphical symbols by applying a parsing process driven by an adjacency grammar. It refers to the symbols and gestures forming the diagram.

Adjacency Grammar to Define Graphical Symbols. An adjacency grammar defines all the symbols and gestures in our framework. Symbols and gestures are formed by tokens. Grammatical productions specify the structure of a symbol, specifying which tokens are forming a symbol and the relations among them.

Following the definition explained above of an adjacency grammar our grammar is defined as a 5-tuple $G = (V_t, V_n, S, P, C)$ where:

- $V_t = \{segment, arc\}$ (See section 2.2).
- $V_n = \{$*QUAD, TRIANGLE, TABLE, CHAIR, CLOSET, SYMBOL33, LIGHTPOINT, PLUG, TELEPHONEPLUG, SWITCH, SYMBOL46, SYMBOL47, SYMBOL50, SYMBOL40, SYMBOL41, SYMBOL27, SELECT, UNDO, DELETE, MOVE, ROTATE*$\}$. According to the entire grammar shown in Fig. 3, a non-terminal symbol represents a graphical symbol or a compounding part.
- $C = \{$*Incident, Adjacent, Intersects, Perpendicular, Parallel, Contains*$\}$. See Fig. 4 for a graphical explanation of such constraints.
- P are the productions of the grammar defined as: *name* $\rightarrow$ $\{$*elements1*,...,$elementsn\}, constraint_1\& \ldots \&constraint_n$
  Where *name* $\in V_n$, $\{$*elements1*,...,*elementsn*$\} \in \{V_n \cup V_t\}$,
  and $\{constraint_1, \ldots, constraint_n\} \in C$. See Fig. 3

In Fig.5, we can see an example of a grammar describing a symbol. With a grammar we can define a symbol directly as the composition of tokens or terminal symbols, or using the decomposition of its parts and defining other non-terminal symbols as a part of the main symbol. As it can be seen in Fig. 5(c), the symbol telephone-plug can be defined in two possible ways. One employing all the tokens forming the symbol.

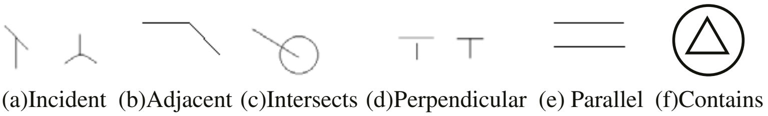
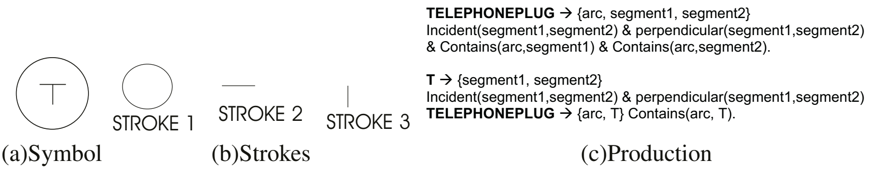

(a)Incident (b)Adjacent (c)Intersects (d)Perpendicular (e) Parallel (f)Contains

Fig. 4. Examples of constraints.



TELEPHONEPLUG → {arc, segment1, segment2}
Incident(segment1,segment2) & perpendicular(segment1,segment2)
& Contains(arc,segment1) & Contains(arc,segment2).

T → {segment1, segment2}
Incident(segment1,segment2) & perpendicular(segment1,segment2)
TELEPHONEPLUG → {arc, T} Contains(arc, T).

(a)Symbol (b)Strokes (c)Production

Fig. 5. Symbol Telephone-Plug (a)Symbol (b)Strokes (c)Productions.

The other using the non-terminal symbol T in the definition of the final symbol. Our grammar to represent symbols can be seen in Fig. 3.

Parsing Process to Recognize Graphical Symbols. Given a sketched symbol, it is recognized by a parsing process guided by the adjacency grammar describing valid symbol classes. There exists several references in the literature of bottom-up parsers for visual languages [9, 10]. These parsers construct the parse tree from the leaves, corresponding to the input primitives, to the root, consisting in the start symbol. In our language, the parser is bottom up and is based on precedences between rules.

The parser process works as follows. Given a set of input tokens, it connects them together with a common root corresponding to a one non-terminal symbol.

Once primitives have been detected in the lexical level a parse tree is hierarchically constructed by iteratively apply the grammar productions from right to left. First, a set of primitives are grouped if their kind and constraints matches the right hand part of a production. Then the left hand symbol is synthesized. This process is iteratively applied until the starting symbol is reached. Tracing the parse tree from the leaves to the root the recognized graphical symbols are identified and also its structure.

The success of the recognition process depends on the good specification of the grammatical rules, and the set of constraints that has been defined.

3 Experimental Results

To test the grammar we have used the CVC online database of symbols[2]. This database is formed by 50 models drawn by 21 persons each, divided into two groups, drawing each person an average of ten instances for 25 symbols. So it results in a database of about 5000 sketched symbols. The acquisition has been done with a digital pen & paper protocol using Logitech io Pen [11]. The purpose of the database is to obtain an important set of symbols that allows to test some different pattern recognition problems. In our case as application framework, we use this database on a sketching architectural application. This project converts a sketched floor plan to a CAD representation consisting of building elements. We have three different kinds of symbols: structures, furniture and utilities. The input of the system can be done by means of a scanner device or a digital pen device. The application also allows the option of interact with the system adding new symbols or by means of gestures. In Fig.6(a), we can see the sketchy input

[2] this database can be obtained contacting the authors of the paper.

of the system, as it can be seen the symbols have distortion, that it makes difficult its recognition. Figure 6(b) shows the output after the recognition process. In Fig. 6(c) and Fig. 6(d), we can see how the user interacts with the system by means of a gesture, *rotate*, and the result of the recognition of the gesture respectively. More details on this application framework can be found in [12].

The performance of our grammar has been tested with the on-line instances of the database. One of the difficulties of these instances is that since they are on-line, we have to take into account a range of tolerance to allow inherent distortion of strokes. In Fig.6(a) we can see a sketch drawn in these framework. Some results of the recognition of symbols with the grammar are shown in the following table:

Table 1. Results with the online database.

	SYMBOL3	SYMBOL5	SYMBOL55	SYMBOL57	SYMBOL33	SYMBOL51	SYMBOL52	SYMBOL46	SYMBOL47	SYMBOL27
PERSON1	100	91	100	82	82	100	80	100	91	100
PERSON2	91	100	100	91	100	100	91	100	100	100
PERSON3	100	100	100	100	85	91	45	100	64	64
PERSON4	100	93	87	73	53	100	27	100	100	36
PERSON5	100	100	100	87	73	100	83	91	100	64
PERSON6	91	82	91	55	45	100	100	91	82	82
BY SYMBOL	97.1	94.6	96.2	83.3	72.4	98.7	74.1	97	89.4	74.2
TOTAL	87.7									

We have chosen instances of the online database from 6 people and apply a grammar that contains the definition of this 10 symbols. The values in the ceils of table1 refer to the success ratio per person in any symbol. The total number of instances is approximately 700 instances. As it can be seen not all the symbols have the same percentage of success and not the same person have the percentage on all the symbols. This is related that there are constraints that are more sensitive to distortion than others. For example, *perpendicular* and *parallel* constraint are more sensitive to distortion than *adjacent* or *intersects*. As said in section 2 the success of the grammar is related to a good specification of the symbols and their constraints.

4 Conclusions

In this paper we have presented an adjacency grammar for on-line parsing diagram-like sketches in digital pen frameworks. The presented approach follows the classical stages of a sketching interface: primitive approximation of input strokes, graphical symbol recognition and interpretation. Two types of symbols have been considered diagram symbols and gestures. Both are called graphical symbols and are formed by line and arc segments. The primitive approximation step extracts from input strokes the set of lines and arc segments considered primitives or tokens. The graphical symbol recognition process parse these tokens using an adjacency grammar which takes into account the possible distortions due to the hand-drawn design.

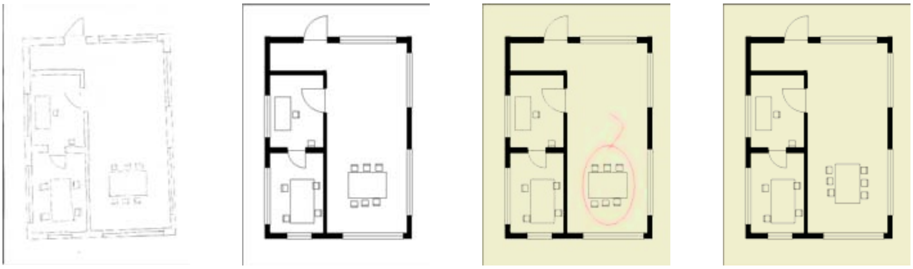

(a)Original (b)Recognized (c)Gesture Rotate (d)Recognized

Fig. 6. Sketching an architectural floor plan.

The performance of the approach have been evaluated using a database of more than 700 on-line sketched symbols getting an average recognition rate of 87.7%. In addition to qualitatively illustrate our work an application to sketch architectural floor plans has been shown.

References

1. Bunke, H.: String grammars for syntactic pattern recognition. In Bunke, H., Sanfeliu, A., eds.: Syntactic and Structural Pattern Recognition. Theory and Applications. World Scientific Publishing Company (1990) 29–54
2. Min, W., Tang, Z., Tang, L.: Using web grammar to recognize dimensions in engineering drawings. Pattern Recognition **26** (1993) 1407–1916
3. Bunke, H.: Attributed programmed graph grammars and their application to schematic diagram interpretation. IEEE Trans. on PAMI **4** (1982) 574–582
4. Fahmy, H., Blonstein, D.: A graph grammar programming style for recognition of music notation. Machine Vision and Applications **6** (1993) 83–99
5. Fahmy, H., Blostein, D.: A survey of graph grammars: Theory and applications. In: Proceedings of 12th. Int. Conf. on Pattern Recognition (a). (1994) 294–298 Jerusalem, Israel.
6. Jorge, J., Glinert, E.: Online parsing of visual languages using adjacency grammars. In: Proceedings of the 11th International IEEE Symposium on Visual Languages. (1995) 250–257
7. Tombre, K., Ah-Soon, C., Dosch, P., Masini, G., Tabonne, S.: Stable and robust vectorization: How to make the right choices. In Chhabra, A., Dori, D., eds.: Graphics Recognition: Recent Advances. Springer-Verlag, Berlin (2000) 3–18 Vol. 1941 of LNCS.
8. Sezgin, T., Stahovich, T., Davis, R.: Sketch based interfaces: Early processing for sketch understanding. In: Proceedings of 2001 Perceptive User Interfaces Workshop (PUI'01), ACM Digital Library (2001) ISBN 1-58113-448-7.
9. Rekers, J., Schurr, A.: Defining and parsing visual languages with layered graph grammars. Journal of Visual Languages and Computing **8** (1997) 27–55
10. Al-Mulhem, M., Ather, M.: Mrg parser for visual languages. Inf. Sci. **131** (2001) 19–46
11. Logitech: IO digital pen (2004) www.logitech.com.
12. Sánchez, G., Valveny, E., Lladós, J., Mas, J., Lozano, N.: A platform to extract knowledge from graphic documents. application to an architectural sketch understanding scenario. In Marinai, S., Dengel, A., eds.: Document Analysis Systems VI. World Scientific (2004) 349–365

Graph Clustering Using Heat Content Invariants

Bai Xiao and Edwin R. Hancock

Department of Computer Science, University of York, York Y01 5DD, UK

Abstract. In this paper, we investigate the use of invariants derived from the heat kernel as a means of clustering graphs. We turn to the heat-content, i.e. the sum of the elements of the heat kernel. The heat content can be expanded as a polynomial in time, and the co-efficients of the polynomial are known to be permutation invariants. We demonstrate how the polynomial co-efficients can be computed from the Laplacian eigensystem. Graph-clustering is performed by applying principal components analysis to vectors constructed from the polynomial co-efficients. We experiment with the resulting algorithm on the COIL database, where it is demonstrated to outperform the use of Laplacian eigenvalues.

1 Introduction

One of the problems that arises in the manipulation of large amounts of graph data is that of embedding graphs in a low dimensional space so that standard machine learning techniques can be used to perform tasks such as clustering. One way of realise this goal is to embed the nodes of a graph on a manifold and to use the geometry of the manifold as a means of characterising the graph. In the mathematics literature, there is a considerable body of work aimed at understanding how graphs can be embedded in manifolds [7]. Broadly speaking there are three ways in which the problem has been addressed. First, the graph can be interpolated by a surface whose genus is determined by the number of nodes, edges and faces of the graph. Second, the graph can be interpolated by a hyperbolic surface which has the same pattern of geodesic (internode) distances as the graph [1]. Third, a manifold can be constructed whose triangulation is the simplicial complex of the graph [12]. A review of methods for efficiently computing distance via embedding is presented in the recent paper of Hjaltason and Samet [5].

The spectrum of the Laplacian matrix has been widely studied in spectral graph theory [4] and has proved to be a versatile mathematical tool that can be put to many practical applications including routing [2], clustering [9] and graph-matching [11]. One of the most important properties of the Laplacian spectrum is its close relationship with the heat equation. The heat equation can be used to specify the flow of information with time across a network or a manifold [10]. According to the heat-equation the time derivative of the kernel is determined by the graph Laplacian. The solution to the heat equation is obtained by exponentiating the Laplacian eigensystem over time. Because the heat kernel encapsulates the way in which information flows through the edges of the graph

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 123–130, 2005.

over time, it is closely related to the path length distribution on the graph. The graph can be viewed as residing on a manifold whose pattern of geodesic distances is characterised by the heat kernel. For short times the heat kernel is determined by the local connectivity or topology of the graph as captured by the Laplacian, while for long-times the solution gauges the global geometry of the manifold on which the graph resides. In a recent paper [13], we have exploited this property and have used heat-kernel embedding for the purposes of graph clustering.

There are a number of different invariants that can be computed from the heat-kernel. Asymptotically for small time, the trace of the heat kernel [4] (or the sum of the Laplacian eigenvalues exponentiated with time) can be expanded as a rational polynomial in time, and the co-efficients of the leading terms in the series are directly related to the geometry of the manifold. For instance, the leading co-efficient is the volume of the manifold, the second co-efficient is related to the Euler characteristic, and the third co-efficient to the Ricci curvature. The zeta-function (i.e. the sum of the eigenvalues raised to a non-integer power) for the Laplacian also contains geometric information. For instance its derivative at the origin is related to the torsion tensor for the manifold. Finally, Colin de Verdiere has shown how to compute geodesic invariants from the Laplacian spectrum [3].

In a recent paper McDonald and Meyers [8] have shown that the heat-content of the heat-kernel is a permutation invariant. The heat content is the sum of the entries of the heat kernel over the nodes of the graph, which may be expanded as a polynomial in time. It is closely related to thre trace of the heat kernel, which is also known to be an invariant. In this paper we show how the co-efficients can be related to the eigenvalues and eigenvectors of the Laplacian. The resulting co-efficients are demonstrated to to outperform the Laplacian spectrum as a means of characterising graph-structure for the purposes of clustering.

2 Heat Kernels on Graphs

In this section, we review the how the heat-kernel is related to the Laplacian eigensystem. To commence, suppose that the graph under study is denoted by $G = (V, E)$ where V is the set of nodes and $E \subseteq V \times V$ is the set of edges. Since we wish to adopt a graph-spectral approach we introduce the adjacency matrix A for the graph where the elements are

$$A(u,v) = \begin{cases} 1 & \text{if } (u,v) \in E \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

We also construct the diagonal degree matrix D, whose elements are given by $D(u,u) = \sum_{v \in V} A(u,v)$. From the degree matrix and the adjacency matrix we construct the Laplacian matrix $L = D - A$, i.e. the degree matrix minus the adjacency matrix. The normalised Laplacian is given by $\hat{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$. The spectral decomposition of the normalised Laplacian matrix is

$$\hat{L} = \Phi \Lambda \Phi^T = \sum_{i=1}^{|V|} \lambda_i \phi_i \phi_i^T \tag{2}$$

where $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_{|V|})$ is the diagonal matrix with the ordered eigenvalues $(0 = \lambda_1 < \lambda_2 \leq \lambda_3 ...)$ as elements and $\Phi = (\phi_1|\phi_2|....|\phi_{|V|})$ is the matrix with the correspondingly ordered eigenvectors as columns. Since $\hat{L}$ is symmetric and positive semi-definite, the eigenvalues of the normalised Laplacian are all positive. The eigenvector ϕ_2 associated with the smallest non-zero eigenvalue λ_2 is referred to as the Fiedler-vector. We are interested in the heat equation associated with the Laplacian, i.e. $\frac{\partial h_t}{\partial t} = -\hat{L} h_t$, where h_t is the heat kernel and t is time. The heat kernel can hence be viewed as describing the flow of information across the edges of the graph with time. The rate of flow is determined by the Laplacian of the graph. The heat kernel, i.e. the solution to the heat equation, is a $|V| \times |V|$ matrix found by exponentiating the Laplacian eigenspectrum, i.e. $h_t = \Phi \exp[-\Lambda t] \Phi^T$. For the nodes u and v of the graph G the resulting element is

$$h_t(u, v) = \sum_{i=1}^{|V|} \exp[-\lambda_i t] \phi_i(u) \phi_i(v) \tag{3}$$

When t tends to zero, then $h_t \simeq I - \hat{L} t$, i.e. the kernel depends on the local connectivity structure or topology of the graph. If, on the other hand, t is large, then $h_t \simeq \exp[-t\lambda_2] \phi_2 \phi_2^T$, where λ_2 is the smallest non-zero eigenvalue and ϕ_2 is the associated eigenvector, i.e. the Fiedler vector. Hence, the large time behavior is governed by the global structure of the graph.

It is interesting to note that the heat kernel is also related to the path length distribution on the graph. If $P_k(u, v)$ is the number of paths of length k between nodes u and v then

$$h_t(u, v) = \exp[-t] \sum_{k=1}^{|V|^2} P_k(u, v) \frac{t^k}{k!} \tag{4}$$

Hence, the heat kernel takes the form of a sum of Poisson distributions over the path-length with time as the parameter. The weights associated with the different components are determined by the associated path-length frequency in the graph. As the path-length k becomes large, the Poisson distributions approach a Gaussian, with mean k and variance k.

The path-length distribution is itself related to the eigenspectrum of the Laplacian. By equating the derivatives of the spectral and path-length forms of the heat kernel it is straightforward to show that

$$P_k(u, v) = \sum_{i=1}^{|V|} (1 - \lambda_i)^k \phi_i(u) \phi_i(v) \tag{5}$$

The geodesic distance between nodes can be found by searching for the smallest value of k for which $P_k(u, v)$ is non zero, i.e. $d_G(u, v) = floor_k P_k(u, v)$.

3 Invariants of the Heat-Kernel

It is well known that the trace of the heat-kernel is invariant to permutations. It is determined by the Laplacian eigenvalues and is given by

$$Z(t) = \sum_{i=1}^{N} \exp[-\lambda_i t] \tag{6}$$

To provide an illustration of the potential utility of the trace-formula, in Figure 1 we show four small graphs with rather different topologies. Figure 2 shows the trace of the heat kernel as a function of t for the different graphs. From the plot it is clear that the curves are distinct and could form the basis of a useful representation to distinguish graphs. For instance, the more bi-partite the graph the more stongly peaked the trace of the heat-kernel at the origin. This is due to the fact the spectal gap, i.e. the size of λ_2, determines the rate of decay of the trace with time, and this in turn is a measure of the degree of separation of the graph into strongly connected subgraphs or "clusters".

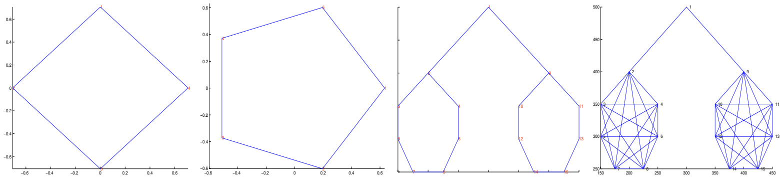


Fig. 1. Four graphs used for heat-kernel trace analysis.

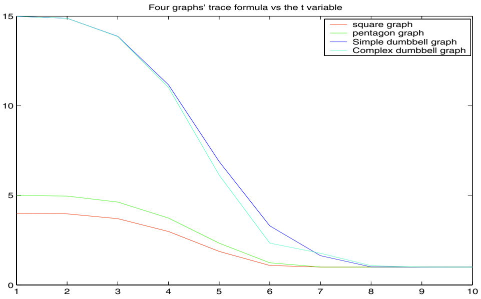


Fig. 2. Heat kernel trace as a function of t for four simple graphs.

Unfortunately, the trace of the heat kernel is limitted use for characterising graphs since for each value of time, it provides only a single scaler attribute. Hence, it must either be sampled with time or a fixed time value selected. However, in a recent paper McDonald and Meyers [8] have shown that the heat-content of the heat-kernel is also an invariant. The heat content is the sum of the entries of the heat kernel over the nodes of the graph and is given by

$$Q(t) = \sum_{u \in V} \sum_{v \in V} h_t(u,v) = \sum_{u \in V} \sum_{v \in V} \sum_{k=1}^{|V|} \exp[-\lambda_k t] \phi_k(u) \phi_k(v) \tag{7}$$

The heat-content can be expanded as a polynomial in time, i.e.

$$Q(t) = \sum_{m=0}^{\infty} q_m t^m \tag{8}$$

By equating the derivatives of the spectral and polynomial forms of the heat content at $t = 0$, the co-efficients are given by

$$q_m = \sum_{i=1}^{|V|} \sum_{u \in V} \sum_{v \in V} \frac{(-\lambda_i)^m}{m!} \phi_i(u) \phi_i(v) \tag{9}$$

In this paper, we will explore the use of the polynomial co-efficients for the purposes of graph-clustering. To do this we construct a vector $\boldsymbol{B} = (q_0, ..., q_5)^T$ from the first six co-efficients of the heat-content polynomial To compare our method with a standard spectral representation we also explore the use of the vector of leading Laplacian eigenvalues $\boldsymbol{B} = (\lambda_2, \lambda_2, ...\lambda_7)^T$ as a feature-vector.

4 Principal Components Analysis

Our aim is to construct a pattern-space for a set of graphs with pattern vectors $\boldsymbol{B}_k$, $k = 1, M$, extracted using heat-content co-efficients. There are a number of ways in which the graph pattern vectors can be analysed. Here, for the sake of simplicity, we use principal components analysis (PCA). We commence by constructing the matrix $\mathbf{S} = [\boldsymbol{B}_1 | \boldsymbol{B}_2 | \ldots | \boldsymbol{B}_k | \ldots | \boldsymbol{B}_M]$ with the graph feature vectors as columns. Next, we compute the covariance matrix for the elements of the feature vectors by taking the matrix product $\mathbf{C} = \mathbf{S}\mathbf{S}^T$. We extract the principal components directions by performing the eigendecomposition $\mathbf{C} = \sum_{i=1}^{M} l_i \boldsymbol{u}_i \boldsymbol{u}_i^T$ on the covariance matrix $\mathbf{C}$, where the l_i are the eigenvalues and the $\boldsymbol{u}_i$ are the eigenvectors. We use the first s leading eigenvectors (3 in practice for visualisation purposes) to represent the graphs extracted from the images. The co-ordinate system of the eigenspace is spanned by the s orthogonal vectors $\boldsymbol{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2, .., \boldsymbol{u}_s)$. The individual graphs represented by the vectors $\mathbf{B}_k, k = 1, 2, \ldots, M$ can be projected onto this eigenspace using the formula $\boldsymbol{\mathcal{B}}_k = \mathbf{U}^T \boldsymbol{B}_k$. Hence each graph G_k is represented by an s-component vector $\boldsymbol{\mathcal{B}}_k$ in the eigenspace.

5 Experiments

We have applied our embedding method to images from the COIL data-base. The data-base contains views of 3D objects under controlled viewer and lighting conditions. For each object in the data-base there are 72 equally spaced views,

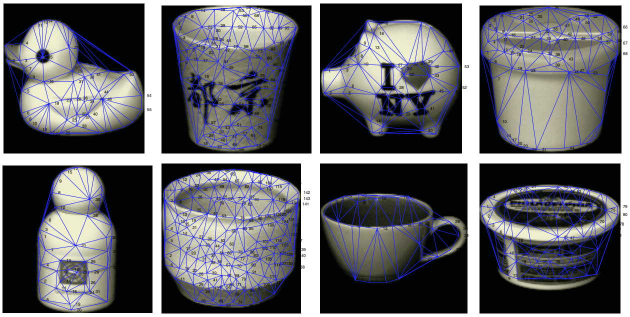


Fig. 3. Eight objects with their Delaunay graphs overlayed.

which are obtained as the camera circumscribes the object. We study the images from eight example objects. A sample view of each object is shown in Figure 3. For each image of each object we extract feature points using the method of [6]. We have extracted graphs from the images by computing the Voronoi tessellations of the feature-points, and constructing the region adjacency graph, i.e. the Delaunay triangulation, of the Voronoi regions. Our embedding procedure has been applied to the resulting graph structures.

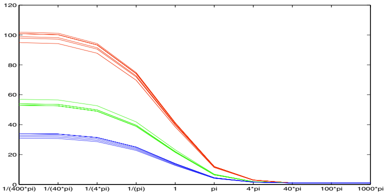


Fig. 4. Heat content as a function of t for 18 COIL graphs.

To commence, we show the heat-content as a function of t for six views of the the second, fifth and seventh objects from the COIL database shown above. From Figure 4 it is clear that objects of the same class trace out curves that are close together. To take this study further, in Figure 5 we plot the six co-efficients q_0, q_1, q_2, q_3, q_4 and q_5 separately as a function of the view number for the eight objects selected from the COIL data-base. The co-efficients are relatively stable with viewpoint. In the left-hand panel of Figure 6 we show the result of performing PCA on the vectors of polynomial co-efficients. For comparison, the right-hand panel in Figure 6 shows the corresponding result when we apply PCA to the vector of leading eigenvalues of the Laplacian matrix $B = (\lambda_2, \lambda_3, ..., \lambda_7)^T$ as the components of a feature vector instead. The main qualitative feature is that the different views of the ten objects are more overlapped than when the heat-content polynomial co-effients are used.

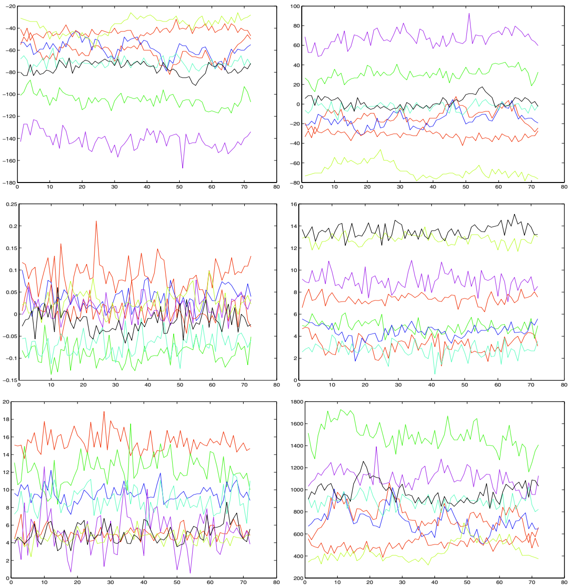


Fig. 5. Individual heat-content co-efficients as a function of view number.

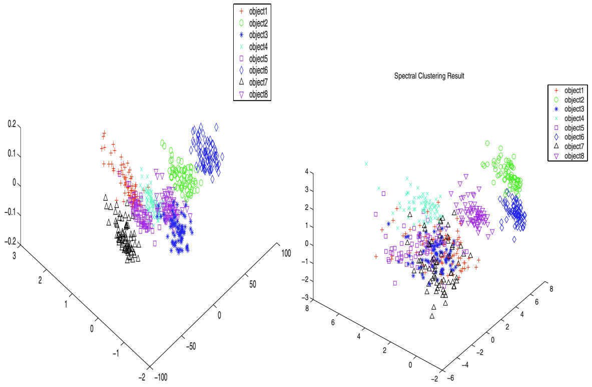


Fig. 6. Applying PCA to the heat-content differential co-efficients (left) and Laplacian spectrum (right).

To investigate the behavior of the two methods in a more quantitative way, we have computed the Rand index for the different objects. The Rand index is defined as $R_I = \frac{C}{C+W}$ where C is the number of ”agreements” and W is the number of ”disagreements” in cluster assignment. The index is hence the fraction of views of a particular class that are closer to an object of the same class than to one of another class. For the heat-content co-efficients, the Rand index is 0.88 while for the Laplacian eigenvalues it is 0.58.

6 Conclusion and Future Work

In this paper we have explored how the use of heat-content can lead to a series of invariants that can be used for the purposes of clustering. There are clearly a number of ways in which the work reported in this paper can be extended. These include the use of features which have a direct geometrical meaning such as the Euler characteristic, the torsion of the mean and Gaussian curvatures of the manifold.

References

1. A.D.Alexandrov and V.A.Zalgaller. Intrinsic geometry of surfaces. *Transl.Math. Monographs*, 15, 1967.
2. J. E. Atkins, E. G. Bowman, and B. Hendrickson. A spectral algorithm for seriation and the consecutive ones problem. *SIAM J. Comput.*, 28:297–310, 1998.
3. Colin de Verdiere. Spectra of graphs. *Math of France*, 4, 1998.
4. F.R.K.Chung. Spectral graph theory. *American Mathematical Society*, 1997.
5. G.R.Hjaltason and H.Samet. Properties of embedding methods for similarity searching in metric spaces. *PAMI*, 25:530–549, 2003.
6. C.G. Harris and M.J. Stephens. A combined corner and edge detector. *Fourth Alvey Vision Conference*, pages 147–151, 1994.
7. N.Linial, E.London, and Y.Rabinovich. The geometry of graphs and some its algorithmic application. *Combinatorica*, 15:215–245, 1995.
8. P.Mcdonald and R.Meyers. Diffusions on graphs, poisson problems and spectral geometry. *Transactions on Amercian Mathematical Society*, 354:5111–5136, 2002.
9. Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE PAMI*, 22:888–905, 2000.
10. S.T.Yau and R.M.Schoen. Differential geometry. *Science Publication*, 1988.
11. S.Umeyama. An eigen decomposition approach to weighted graph matching problems. *IEEE PAMI*, 10:695–703, 1988.
12. S.Weinberger. Review of algebraic l-theory and topological manifolds by a.ranicki. *BAMS*, 33:93–99, 1996.
13. X.Bai and E.R.Hancock. Heat kernels, manifolds and graph embedding. *IAPR Workshop on Structural,Syntactic, and Statistical Pattern Recognition, Lecture Notes in Computer Science 3138*, pages 198–206, 2004.

Matching Attributed Graphs: 2nd-Order Probabilities for Pruning the Search Tree

Francesc Serratosa[1] and Alberto Sanfeliu[2]

[1] Universitat Rovira i Virgili, Dept. d'Enginyeria Informàtica i Matemàtiques, Spain
Francesc.Serratosa@urv.net
http://www.etse.urv.es/~fserrato
[2] Universitat Politècnica de Catalunya, Institut de Robòtica i Informàtica Industrial, Spain
sanfeliu@iri.upc.es

Abstract. A branch-and-bound algorithm for matching Attributed Graphs (AGs) with Second-Order Random Graphs (SORGs) is presented. We show that the search space explored by this algorithm is drastically reduced by using the information of the 2nd-order joint probabilities of vertices of the SORGs. A SORG is a model graph, described elsewhere, that contains 1st and 2nd-order order probabilities of attribute relations between elements for representing a set of AGs compactly. In this work, we have applied SORGs and the reported algorithm to the recognition of real-life objects on images and the results show that the use of 2nd-order relations between vertices is not only useful to decrease the run time but also to increase the correct classification ratio.

1 Introduction

A Second Order Random Graph (SORG) is a model graph introduced by the authors that contains 1st-order and 2nd-order probabilities of attributes to describe a set of Attributed Graphs (AGs) [1,2]. Let us consider, as an example, the 3D-object modelling and recognition problem. The basic idea is that only a single SORG is synthesised from the AGs that represent several views of a 3D-object. Therefore, in the recognition process, only one comparison is needed between each object model represented by a SORG and the unclassified object (view of a 3D-object) represented by an AG.

SORGs can be seen as a generalisation of FDGs [3,4] and First-Order Random Graphs [5,6]. Moreover, Bunke [7] presented a model of sets of graphs, called *network of models,* in which all the graphs are pre-processed generating a symbolic data structure. In the SORGs, to deal with the 1st-order and 2nd-order probabilities, there is a random variable α_i (or β_i) associated with each vertex ω_i (or arc ε_i, respectively), which represents the attribute information of the corresponding graph elements in the set of AGs. A random variable has a probability density function p_i defined over the same attribute domain of the AGs, including a null value Φ that denotes the non-instantiation of an SORG graph element in an AG.

The distance measure between an AG and a SORG was proposed in [2] for error-tolerant graph matching. Here, we present a branch-and-bound algorithm which computes exactly this distance measure. This algorithm uses the 2nd-order probabilities in the SORG to prune the search tree.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 131–138, 2005.

2 Formal Definitions of Random-Graph Representation

Definition 1: Attributed Graph (AG). Let Δ_v and Δ_e denote the domains of possible values for attributed vertices and arcs, respectively. These domains are assumed to include a special value Φ that represents a *null value* of a vertex or arc. An AG G over (Δ_v,Δ_e) is defined to be a four-tuple $G=(\Sigma_v,\Sigma_e,\gamma_v,\gamma_e)$, where $\Sigma_v=\{v_k \mid k=1,...,n\}$ is a set of vertices (or nodes), $\Sigma_e=\{e_{ij} \mid i,j\in\{1,...,n\}, i\neq j\}$ is a set of arcs (or edges), and the mappings $\gamma_v:\Sigma_v\rightarrow\Delta_v$ and $\gamma_e:\Sigma_e\rightarrow\Delta_e$ assign attribute values to vertices and arcs.

Definition 2: Random Graph (RG). Let Ω_v and Ω_e be two sets of random variables with values in Δ_v (random vertices) and in Δ_e (random arcs), respectively. A *random-graph structure R* over (Δ_v,Δ_e) is defined to be a tuple $(\Sigma_v,\Sigma_e,\gamma_v,\gamma_e,P)$, where $\Sigma_v=\{\omega_k \mid k=1,...,n\}$ is a set of vertices, $\Sigma_e=\{\varepsilon_{ij} \mid i,j\in\{1,...,n\}, i\neq j\}$ is a set of arcs, the mapping $\gamma_v:\Sigma_v\rightarrow\Omega_v$ associates each vertex $\omega_k\in\Sigma_v$ with a random variable $\alpha_k=\gamma_v(\omega_k)$ with values in Δ_v, and $\gamma_e:\Sigma_e\rightarrow\Omega_e$ associates each arc $\varepsilon_{ij}\in\Sigma_e$ with a random variable $\beta_k=\gamma_e(\varepsilon_{ij})$ with values in Δ_e. And, finally, P is a joint probability distribution $P(\alpha_1,\ldots,\alpha_n,\beta_1,\ldots,\beta_m)$ of all the random vertices $\{\alpha_i \mid \alpha_i=\gamma_\omega(\omega_i), 1\leq i\leq n\}$ and random arcs $\{\beta_j \mid \beta_j=\gamma_\varepsilon(\varepsilon_{kl}), 1\leq j\leq m\}$.

Definition 3: Probability of a RG instantiation. Given an AG G and a RG R, the joint probability of random vertices and arcs is defined over an instantiation that produces G, and such instantiation is associated with a structural isomorphism $\mu: G'\rightarrow R$, where G' is the extension of G to the order of R. G' represents the same object than G but some vertices or arcs have been added with the null value Φ to be μ bijective. Let G be oriented with respect to R by the structurally coherent isomorphism μ; for each vertex ω_i in R, let $\mathbf{a}_i=\gamma_v(\mu^{-1}(\omega_i))$ be the corresponding attribute value in G', and similarly, for each arc ε_{kl} in R (associated with random variable β_j) let $\mathbf{b}_j=\gamma_e(\mu^{-1}(\varepsilon_{kl}))$ be the corresponding attribute value in G'. Then the *probability of G according to* (or given by) *the orientation* μ, denoted by $P_R(G|\mu)$, is defined as

$$P_R(G|\mu)=\Pr\left(\wedge_{i=1}^{n}(\alpha_i=\mathbf{a}_i)\wedge\wedge_{j=1}^{m}(\beta_j=\mathbf{b}_j)\right)=p(\mathbf{a}_1,\ldots,\mathbf{a}_n,\mathbf{b}_1,\ldots,\mathbf{b}_m) \quad (1)$$

We define $\mathbf{d}_i$ to represent a vertex or arc attribute value ($\mathbf{a}_i$ or $\mathbf{b}_i$). Thus, if s is the number of vertices and arcs, $s=m+n$, eq. (1) can be rewritten as,

$$P_R(G|\mu)=p(\mathbf{d}_1,\ldots,\mathbf{d}_s) \quad (2)$$

3 Second-Order Random-Graph Representation

If we want to represent the cluster of AGs by a RG, it is impractical to consider the high order probability distribution defined in the RGs $P(\alpha_1,\ldots,\alpha_n,\beta_1,\ldots,\beta_m)$ (definition 2), where all components and their relations in the structural patterns are taken jointly due to time and space costs. For this reason, some other more practical approaches have been presented that propose different approximations [1,4,5,6]. All of them take into account in some manner the incidence relations between attributed vertices and arcs, i.e. assume some sort of dependence of an arc on its connecting vertices. Also, a common ordering (or labelling) scheme is needed that relates vertices and arcs of all the involved AGs, which is obtained through an optimal graph mapping process called synthesis of the random graph representation. We showed in [1] that all the approximations in the literature of the joint probability of an instantiation of the random elements in a RG (eq. 1) can be described in a general form as follows:

$$P_R(G|\mu)=p(\mathbf{a}_1,,,\mathbf{a}_n,\mathbf{b}_1,,,\mathbf{b}_m)=\prod_{i=1}^{n}p_i(\mathbf{a}_i)\prod_{i=1}^{m}p_i(\mathbf{b}_i)\prod_{i=1}^{n-1}\prod_{j=i+1}^{n}r_{ij}(\mathbf{a}_i,\mathbf{a}_j)\prod_{i=1}^{n}\prod_{j=1}^{m}r_{ij}(\mathbf{a}_i,\mathbf{b}_j)\prod_{i=1}^{m-1}\prod_{j=i+1}^{m}r_{ij}(\mathbf{b}_i,\mathbf{b}_j) \quad (3)$$

where p_i are the marginal probabilities of the s random elements γ_i, (vertices or arcs) and r_{ij} are the Peleg compatibility coefficients [9] that take into account both the marginal and 2nd-order joint probabilities of random vertices and arcs.

According to eq. (2), we can generalise the joint probability as,

$$P_R(G|\mu)=p(\mathbf{d}_1,,,\mathbf{d}_s)=\prod_{i=1}^{s}p_i(\mathbf{d}_i)\prod_{i=1}^{s}\prod_{j=i+1}^{s}r_{ij}(\mathbf{d}_i,\mathbf{d}_j) \quad (4)$$

and define the Peleg coefficient,

$$r_{ij}(\mathbf{d}_i,\mathbf{d}_j)=\frac{p_{ij}(\mathbf{d}_i,\mathbf{d}_j)}{p_i(\mathbf{d}_i)p_j(\mathbf{d}_j)} \quad (5)$$

The Peleg coefficient, with a non-negative range, is related to the "degree" of dependence between two random variables. If they are independent, the joint probability, p_{ij}, is defined as the product of the marginal ones, thus, $r_{ij} = 1$ (or a value close to 1 if the probability functions are estimated). If one of the marginal probabilities is null, the joint probability is also null. In this case, the indecisiveness *0/0* is solved as 1, since this do not affect the global joint probability, which is null.

4 Distance Measure Between AGs and SORGs

The distance measure presented in this section provides a quantitative value of the match between an AG *G* (data graph) and a SORG *S* (model graph) similar to the one presented in [2]. It is related to the probability of G according to the labelling function $\mu : G \to S$, denoted $P(G|\mu)$ in eq. (4). We may attempt to minimise a *global cost* measure *C* of the morphism μ in the set *H* of allowable configurations, by taking the cost as a monotonic decreasing function of the conditional probability of the data graph given the labelling function, $C = f(P(G|\mu))$. With some steps depicted in [2] we arrive to the final expression

$$C(G|\mu)=-(s-2)\sum_{i=1}^{s}C_i^1(\mathbf{d}_i)+\sum_{i=1}^{s-1}\sum_{j=i+1}^{s}C_{i,j}^2(\mathbf{d}_i,\mathbf{d}_j) \quad (6)$$

where first-order and second order costs are given by

$$C_i^1(\mathbf{d}_i) = Cost(p_i(\mathbf{d}_i)) \qquad C_{i,j}^2(\mathbf{d}_i,\mathbf{d}_j) = Cost(p_{i,j}(\mathbf{d}_i,\mathbf{d}_j)) \tag{7}$$

and the function *Cost(Pr)* yields a bounded normalized cost value between 0 and 1 depending on the negative logarithm of a given probability *Pr* and parameterised by a positive constant $K_{pr} \in [0,1]$, which is a threshold on low probabilities that is introduced to avoid the case ln(0), which would give negative infinity. This is,

$$Cost(\mathrm{Pr}) = \begin{cases} \dfrac{-\ln(\mathrm{Pr})}{-\ln(K_{\mathrm{Pr}})} & \text{if } \mathrm{Pr} \geq K_{\mathrm{Pr}} \\ 1 & \text{otherwise} \end{cases} \tag{8}$$

Once a cost measure C is defined, a distance measure between an AG and a SORG and the optimal labelling μ^* are defined respectively as

$$d = \min_{\mu \in H} \{C(G|\mu)\} \qquad \text{and} \qquad \mu^* = \arg\min_{\mu \in H} \{C(G|\mu)\} \tag{9}$$

5 Algorithm for Computing the Distance Measure

The distance and the optimal morphism between an AG G and an SORG F are calculated by an algorithm for error-tolerant graph matching. Our approach is based on a tree search by A* algorithm, where the search space is reduced by a *branch and bound* technique. The algorithm searches a tree where the nodes represent possible mappings between vertices of both graphs and branches represent combinations of pairs of graph vertices that satisfy the labelling constraints. Hence, the paths from the root to the leaves represent allowed labellings f.

The distance measure has been theoretically defined such that both graphs are extended with null elements to have the same number of elements and to be complete. Nevertheless, in practice, our algorithm only needs the SORG to be extended with one null vertex, because the different permutations of the null vertices are regarded as equivalent labellings. Thus, the AG spurious vertices are possibly matched with this unique null SORG vertex (ω_Φ) and hence the mapping is not forced to be injective. On the other hand, the SORG graph elements that remain unmatched when arriving at a leaf are considered to be matched with null AG vertices v_Φ or null AG arcs e_Φ. Consequently, a final cost of deleting these elements may be added to the cost of the labelling in the leaves of the search tree. Nevertheless, if a sub-graph isomorphism from an AG to a SORG is looked for, then it is not needed to match all the SORG vertices with an AG vertex (null or not) and this deleting cost has not to be computed. This is the case of our application in Section 5.

In general, solving a branch and bound problem requires a *branch evaluation function* and a *global evaluation function*. The former assigns a cost to the branch incident to a node N of the tree, which is the cost of the new match (or pair) appended. The latter is used to guide the search at a node N and refers to the cost of the best complete path through N (i.e. including the pairs of vertices already matched when arriving at N). The cost of a labelling f is given by the value of the *global evaluation function* in a leaf of the search tree.

Each node *N* of the search tree at level p>0 is described by a collection of pairs of vertices of the graphs, $N = \{(v_i, \omega_{q_i})\}$, where $i = 1,2,...,p$ correspond to the indices of the vertices v_i in the AG and q_i are the distinct indices of the vertices ω_{qi} in the SORG such that $f(v_i) = \omega_{q_i}$. Moreover, we define the sets $N_v = \{v_1, v_2, ..., v_p\}$ and $N_\omega = \{\omega_{q_1}, \omega_{q_2}, ..., \omega_{q_p}\}$ of vertices that have already been matched between both graphs, and also, the sets $M_v = \{v_{p+1}, v_{p+2}, ..., v_n\}$ and $M_\omega = \{\omega_j \mid \omega_j \notin N_\omega\}$ of vertices that have not been matched yet. Assume that $N = \{(v_1, \omega_{q_1}), (v_2, \omega_{q_2}), ..., (v_p, \omega_{q_p})\}$ indicates the unique path from the root to a tree node *N* and $T = \{(v_1, \omega_{q_1}), (v_2, \omega_{q_2}), ..., (v_n, \omega_{q_n})\}$ indicates the unique path from the root to a leaf *T*. The vertices of M_ω in each node *N* are explored using the order imposed by the costs $C^1(v_p, \omega_j)$ being $\omega_j \in M_\omega$.

The *branch evaluation function K* depends on the cost of the new match between vertices, the cost of all the arcs related to these two vertices that involve vertices from N_v and N_ω and the 2nd-order costs referring to these same vertices. Thus, the cost assigned to the branch incident to *N* is given by

$$K(v_p, \omega_{q_p}) = -(p-3)\left(C^1(v_p, \omega_{q_p}) + \sum_{s=1}^{p-1} C^1(e_{ps}, \varepsilon_{q_p q_s}) + C^1(e_{sp}, \varepsilon_{q_s q_p})\right) + \sum_{s=1}^{p-1} C^2(v_p, v_s, \omega_{q_p}, \omega_{q_s}) \tag{10}$$

The *global evaluation function* $l^*(N)$ at a node *N* of level *p* is defined as the cost $g^*(N)$ of an optimal path from the root to the node *N* plus the cost $h^*(N)$ of an optimal path from the node *N* to a leaf $T = \{(v_i, \omega_{q_i}) \mid i = 1,2,...,n\}$ constrained to be reached through the node *N*:

$$l^*(N) = g^*(N) + h^*(N) \quad \text{where} \quad g^*(N) = \sum_{i=1}^{p} K(v_i, \omega_{q_i}) \quad \text{and} \quad h^*(N) = \min_t \sum_{i=p+1}^{n} K(v_i, \omega_{q_i}) \tag{11}$$

where *t* denotes a feasible path from *N* to *T*.

On the other hand, the global evaluation function $l^*(N)$ is unknown in an inner node *N*, since $h^*(N)$ is unknown and can only be approximated by a consistent lower-bounded estimate.

For that purpose, let $K'(v_i, \omega_j)$ be the cost of adding a pair of vertices to *N*, where $v_i \in M_v$ and $\omega_j \in M_\omega$, defined as

$$K'(v_i, \omega_j) = -(p-3)\left(C^1(v_i, \omega_j) + \sum_{s=1}^{p} C^1(e_{is}, \varepsilon_{jq_s}) + C^1(e_{si}, \varepsilon_{q_s j})\right) + \sum_{s=1}^{p} C^2(v_i, v_s, \omega_j, \omega_{q_s}) \tag{12}$$

Then, for each unmatched vertex $v_i \in M_v$, a corresponding vertex $\omega_j \in M_\omega$ can be associated such that the cost $K'(v_i, \omega_j)$ is minimised.

Finally, the *heuristic function* $l(N)$ that estimates $l^*(N)$ in a node *N* is given by $l(N) = g^*(N) + h(N)$ where $h(N) = \sum_{i=p+1}^{n} \min_{\forall \omega_j \in M_\omega} \{K'(v_i, \omega_j)\}$ is a consistent lower bounded estimate of $h^*(N)$.

The algorithm to compute the distance measure d and the corresponding optimal labelling f_{opt} between a given AG and a given FDG only invokes the following recursive procedure TreeSearch at the root node.

```
Procedure TreeSearch($G,F,f,g^*,v_i,d_f,f_{opt}$)
Input parameters: $G$ and $F$: An AG and a SORG
$f$: Optimal path (labelling) from the root to the current node
$g^*$: Minimum value from the root to the current node
$v_i$: Current AG vertex to be matched
Let $W$ be a sequence of $w_j$ of $F$ ordered by $C^1(v_i,w_j)$
For each vertex $w_j$ in $W$ not used yet in $f$ or $w_\Phi$ do
$K$:=Branch-Evaluation-Function($G,F,f,v_i,w_j$)
$h$:=Bound-Estimate-Function($G,F,fU\{f(v_i)=w_j\}$)
$l:=g^*+K+h$ ;   {Heuristic function of $l^*$}
    If $l<d_f$ then {partial cost < best distance}
        If $I<n$ then   {some vertex still not matched}
          TreeSearch($G,F, fU\{f(v_i)=w_j\},g^*+K,v_{i+1},d_f,f_{opt}$)
        Else  {all AG vertices have been matched}
              $d_f:=l$;
                  $f_{opt}:=fU\{f(v_i)=w_j\}$
End-procedure
```

6 Recognition of Real-Life Objects on Images

We present a real application to recognise coloured objects using 2D images. Images were extracted from the database COIL-100 from Columbia University (www.cs.columbia.edu/CAVE/research/ softlib/coil-100.html). It is composed by 100 isolated objects and for each object there are 72 views (one view each 5 degrees). Figure 1 shows some objects at angle 100 and their segmented images with the adjacency graphs. These graphs have from 6 to 18 vertices and the average number is 10. The test set was composed by 36 views per object (taken at the angles 0, 10, 20 and so on), whereas the reference set was composed by the 36 remaining views (taken at the angles 5, 15, 25 and so on). We compared SORGs to 3 other classifiers. The probabilistic models First-Order Random Graphs (FORGs) [6], Function-Described Graphs (FDGs) [4] and the Nearest-Neighbour classifier (AG-AG) with the edit-operation distance between graphs as measure of similarity.

We made 6 different experiments in which the number of clusters that represents each 3D-object varied. If the 3D-object was represented by only one cluster, the 36 AGs from the reference set that represent the 3D-object were used to synthesise the SORGs, FORGs or FDGs. If it was represented by 2 clusters, the 18 first and consecutive AGs from the reference set were used to synthesise one of the SORGs, FORGs or FDGs and the other 18 AGs were used to synthesise the other ones. A

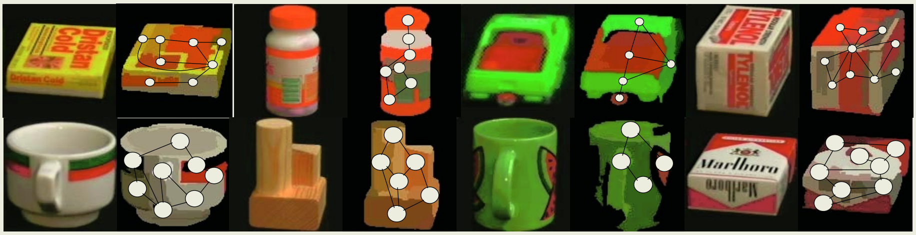


Fig. 1. Some objects at angle 100 and the segmented images with the AGs

similar method was used for the other experiments with 3, 4, 6 and 9 clusters per 3D-object.

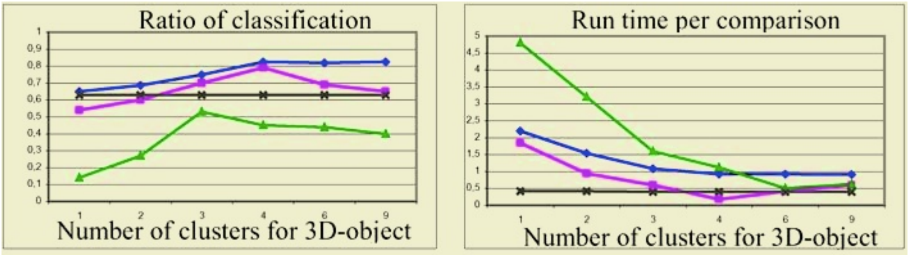


Fig. 2. (a) Ratio of recognition correctness (b) run time spent in the classification. SORG:—◆—; FDG:—■—; FORG:—▲—; AG-AG:—✕—

Figure 2.a shows the ratio of correctness of the four classifiers varying the number of clusters per each object. When objects are represented by only 1 or 2 clusters, there are too much spurious regions (produced in the segmentation process) to keep the structural and semantic knowledge of the object. For this reason, different regions or faces (or vertices in the AGs) of different views (that is, AGs) are considered to be the same face (or vertex in the AGs). The best result appears when each object is represented by 3 or 4 clusters, that is, each cluster represents 90 degrees of the 3D-object. When objects are represented by 9 clusters, each cluster represents 40 degrees of the 3D-object and 4 AGs per cluster, there is poor probabilistic knowledge and therefore the distance costs on the vertices and arcs are coarse. Figure 2.b shows the average run time spent to compute the classification. When the number of clusters per object decreases, the number of total comparisons also decreases but the time spent to compute the distance increases since the structures that represent the clusters (SORGs, FORGs or FDGs) are bigger.

When the best classification is reached, FDGs have less run-time than SORGs but lower recognition ratio. This is due to the fact that the algorithm to compute the distance in the FDG classifier prunes harder the search tree than the SORGs since it uses a qualitative information of the 2nd-order relation [3]. Therefore, the time spent to search the best labelling decreases but the optimal one may not be found.

7 Conclusions and Future Work

We have presented an algorithm to compute the distance measure between AGs and SORGs. It is based on a well known algorithm that uses the branch-and-bound technique and the distance between vertices and arcs as the *heuristic function* to prune the search tree. Due to the fact that SORGs keep 2nd-order probabilities between vertices, we incorporate this knowledge into the *heuristic function* to reduce harder the space explored by the algorithm.

The experimental results show that, in the FDG classifier, the use of the antagonism relations between vertices, is useful not only to decrease the run time of the matching algorithm but also to increase the recognition ratio (thanks to a better modelling of the set of objects). Nevertheless, if the 2nd-order probabilities are kept in the model as in the case of the SORGs, the classification ratio increases but the run time also increases.

We are defining a matching algorithm that computes a sub-optimal distance between AGs and SORGs in polynomial cost. It is based on the distance between cliques that uses 2nd-ordre probabilities between the external vertices of both cliques. This distance will be useful to select some SORG candidates. Thus, the distance algorithm explained in this article with exponential cost will be only applied to this few candidates.

References

1. Sanfeliu, F. Serratosa & R. Alquézar, “Second-Order Random Graphs for modeling sets of Attributed Graphs and their application to object learning and recognition”*, International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 18, No. 3, pp: 375-396, 2004.
2. F. Serratosa & A. Sanfeliu, “Distance measures between Attributed Graphs and Second-order Random Graphs”, *Proc. Syntactic and Structural Pattern Recognition, SSPR’2004, LNCS 3138*, pp: 1135-1144, 2004.
3. F. Serratosa, R. Alquézar y A. Sanfeliu, “Synthesis of function-described graphs and clustering of attributed graphs”, *Int. Journ. of Pattern Recognition and Artificial Intelligence*, 16, (6), pp.621-655, 2002.
4. F. Serratosa, R. Alquézar y A. Sanfeliu, “Function-Described Graphs for modeling objects represented by attributed graphs”, *Pattern Recognition*, 36 (3), pp. 781-798, 2003.
5. K. Sengupta and K. Boyer, “Organizing large structural modelbases”, *PAMI*, vol. 17, pp.321-332, 1995
6. A.K.C. Wong and M. You, “Entropy and distance of random graphs with application to structural pattern recognition”, *IEEE PAMI.*, vol. 7, pp. 599-609, 1985.
7. H. Bunke, “Error-tolerant graph matching: a formal framework and algorithms”. *SSPR’98 & SPR’98,* Sydney, Australia, Springer LNCS-1451,pp.1-14, 1998.
8. A.K.C. Wong, M. You and S.C. Chan, “An algorithm for graph optimal monomorphism”, *IEEE Trans. on Systems, Man and Cyber*, vol. 20, pp. 628-636, 1990.
9. S. Peleg and A. Rosenfeld, “Determining compatibility coefficients for curve enchancement relaxation processes”, *IEEE Transactions on Systems, Man and Cybernetics*, vol. 8, pp. 548-555, 1978.

Synthesis of Median Spectral Graph

Miquel Ferrer[1], Francesc Serratosa[1], and Alberto Sanfeliu[2]

[1] Universitat Rovira I Virgili, Dept. d'Enginyeria Informàtica i Matemàtiques, Spain
{miquel.ferrer,francesc.serratosa}@urv.net
[2] Universitat Politècnica de Catalunya, Institut de Robòtica i Informàtica Industrial, Spain
anfeliu@iri.upc.es

Abstract. In pattern recognition, median computation is an important technique for capturing the important information of a given set of patterns but it has the main drawback of its exponential complexity. Moreover, the Spectral Graph techniques can be used for the fast computation of the approximate graph matching error, with a considerably reduced execution complexity. In this paper, we merge both methods to define the Median Spectral Graphs. With the use of the Spectral Graph theories, we find good approximations of median graph. Experiments on randomly generated graphs demonstrate that this method works well and it is robust against noise.

1 Introduction

Attributed Graphs (AGs) has been used to solve computer vision problems for decades and in many applications. Some examples include recognition of graphical symbols, character recognition, shape analysis, 3D-object recognition and video and image database indexing. In these applications, AGs represent both unclassified objects (unknown input patterns) and prototypes. Moreover, these AGs are typically used in the context of nearest-neighbour classification. That is, an unknown input pattern is compared with a number of prototypes stored in the database. The unknown input is then assigned to the same class as the most similar prototype.

Nevertheless, the main drawback of representing the data and prototypes by AGs is the computational complexity of comparing two AGs. The time required by any of the optimal algorithms may in the worst case become exponential in the size of the AGs. The approximate algorithms, on the other hand, have only polynomial time complexity, but do not guarantee to find the optimal solution.

Moreover, in some applications, the classes of objects are represented explicitly by a set of prototypes which means that a huge amount of model AGs must be matched with the input AG and so the conventional error-tolerant graph matching algorithms must be applied to each model-input pair sequentially. As a consequence, the total computational cost is linearly dependent on the size of the database of model graphs and exponential (or polynomial in subgraph methods) with the size of the AGs. For applications dealing with large databases, this may be prohibitive.

To alleviate these problems, some attempts have been made to try to reduce the computational time of matching the unknown input patterns to the whole set of models from the database. Assuming that the AGs that represent a cluster or class are not completely dissimilar in the database, only one structural model is defined from the AGs that represent the cluster, and thus, only one comparison is needed for each cluster.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 139–146, 2005.

There are two different methodologies to represent the cluster in the literature depending on whether they keep probabilistic information in the structure that represent the cluster of AGs or not. In the probabilistic methods, the models (clusters) are described in the most general case through a joint probability space of random variables ranging over graph vertices and arcs. They are the union of the AGs in the cluster, according to some synthesis process, together with its associated probability distribution [1,2,3]. In the non probabilistic methods, clusters are represented by an AG (which might not be in the cluster) or they are represented by a network of models [4].

Spectral graph theory is concerned with understanding how the structural properties of graphs can be characterised using the eigenvectors of the adjacency matrix of the AGs or the Covariance matrix [5]. Although spectral methods have been used to address the segmentation or correspondence matching problems, there has been less work on using spectral characteristics to perform pattern analysis on AGs. First, an approximate solution to the graph matching problem was presented in [6,7,8] for both undirected and directed AGs based on the eigendecomposition of the adjacency matrix of both graphs. The method was restricted to AGs with only one positive attribute on the nodes and arcs. Recently, AGs with complex numbers as attributes on the nodes or arcs were allowed in the method presented in [9,10], rather than purely real entries.

Given a set of AGs, the median is defined as the AG that has the smallest sum of the distances to all AGs in the set [11]. We can distinguish between set median and generalised median graphs. The difference lies in the space of AGs where the respective median is searched for (formal definitions in section 5).

In this paper, we first define a method to find a sub-optimal labelling between AG vertices based on the correlation between the modal matrices obtained from the adjacency matrices of both AGs (section 4). Moreover, we introduce the novel concepts of set and generalised-median spectral graphs (section 5). While the computation for the set-median spectral graphs is exponential in the size of the input graphs, but polynomially bounded by the number of those graphs, the complexity of computing generalised-median spectral graphs is exponential in both the number of input graphs and their size. For this reason, and with the aim of reducing the exponential complexities, we develop an incremental algorithm in section 6 to synthesise an approximation of the generalised-spectral graph in polynomial complexity respect the number of AGs. Experiments conducted on median spectral graphs in section 7 demonstrate the advantage of this representation and the ability of our synthesis method to find approximate generalised-median spectral graphs.

2 Formal Definitions of Attributed Graphs

An ***attributed graph*** *G* over the domain of the attribute vertices and arcs (Δ_v, Δ_e) with an underlying graph structure $H = (\Sigma_v, \Sigma_e)$, where $\Sigma_v = \{v_k \mid k = 1,...,n\}$ is a set of vertices (or nodes) and $\Sigma_e = \{e_{ij} \mid i, j \in \{1,...,n\}, i \neq j\}$ is a set of arcs, is defined to be a pair (V, E) where $V = (\Sigma_v, \gamma_v)$ is an *attributed vertex set* and $E = (\Sigma_e, \gamma_e)$ is an *attributed arc set*. The mappings $\gamma_v : \Sigma_v \rightarrow \Delta_\omega$ and $\gamma_e : \Sigma_e \rightarrow \Delta_\varepsilon$ assign attribute values to vertices and arcs, respectively, where $\Delta_\varepsilon = \Delta_e \cup \{\Phi\}$ and $\Delta_\omega = \Delta_v \cup \{\Phi\}$. Φ ($\Phi \notin \Delta_v$ and $\Phi \notin \Delta_e$.) represents a *null value* of a graph element.

An ***adjacency matrix*** *A* of an attributed graph *G* of order *n* is a *n X n* matrix whose element with row index *i* and column index *j* is:

$$A(i,j) = \begin{cases} \gamma_e(e_{ij}) & if \quad e_{ij} \in \Sigma_e \\ \Phi & otherwise \end{cases}$$

3 Spectral Decomposition of Matrices

Spectral decomposition of matrices: The spectral decomposition of matrices is obtained as follows. From matrix *A* to hand, we can calculate the eigenvalues λ = (λ$_1$, λ$_2$, …,λ$_n$) by solving the equation $|A - \lambda I| = 0$. Moreover, the modal matrix (also called eigenvector matrices) $U = (u_1| \ u_2| \ldots \ u_n)$, composed by the eigenvectors associated to the eigenvalues λ, is obtained by solving the system of equations $Au_w = \lambda u_w$, were *w* is the eigenmode index and the order of the eigenvectors is decided according to the decreasing magnitude of the eigenvalues, i.e. $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. We emphasise that in the case that the initial matrix is not symmetric, the elements of the modal matrix are complex numbers but in the case that the initial matrix is symmetric, the elements of the modal matrix are real numbers. The original matrix *A* can be recovered by its eigenvectors and eigenvalues, $A = U \ diag(\lambda) \ U^T$. See [12] for more details.

Correlation between matrices: Given a pair of matrices $A = (a_1| \ a_2| \ldots \ a_n)$, and $B = (b_1| \ b_2| \ldots \ b_n)$ of *nXn* rows and columns, the correlation Γ between them is defined as,

$$\Gamma(A,B) = \max_{\gamma} \sum_{i=1}^{n} a_i b_{\gamma(i)}^T \tag{1}$$

4 Error-Tolerant Graph Matching

From this section to the rest of the paper we consider that the domain of the vertices and arcs, Δ_e, is the set of the non-negative numbers. The null attribute, Φ, is represented by zero. And also, nodes have no attributes, that is, $\Delta_v = \{\Phi\}$. With this conditions, the adjacency matrix totally characterises the AGs and it is composed by non-negative numbers.

4.1 Distance Between Attributed Graphs Given a Labelling

Let $G^1 = (V^1, E^1)$ and $G^2 = (V^2, E^2)$ be two AGs with *n* nodes. A *global cost* C_f can be associated with each structurally correct labelling between vertices of both graphs $f^{1,2} : \Sigma_v^1 \rightarrow \Sigma_v^2$, and the distance measure between them is defined as the minimum of all such costs [13]:

$$d(G^1, G^2) = \min_{f^{1,2}} \left\{ C_{f^{1,2}}(G^1, G^2) \right\} \tag{2}$$

We call the *optimal labelling* as the labelling that gives the minimum cost, that is, the one used to compute the distance. Finally, the global cost $C_{f^{1,2}}$ is defined as the difference of the attribute values as follows,

$$C_{f^{1,2}}\left(G^1,G^2\right)=\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\gamma_e^1\left(e_{ij}^1\right)-\gamma_e^2\left(e^2_{\arg\left(f^{1,2}\left(v_i^1\right)\right),\arg\left(f^{1,2}\left(v_j^1\right)\right)}\right)\right)^2 \tag{3}$$

If A_{G^1} and A_{G^2} are the adjacency matrices of the AGs G^1 and G^2, the global cost $C_{f^{1,2}}$ defined in (2) can be reformulated as follows using a permutation matrix *P* [6]:

$$C_{f^{1,2}}=\left\|PA_{G^1}P^T-A_{G^2}\right\|^2 \tag{4}$$

where *P* represents the isomorphism $f^{1,2}$, that is,

$$P(i,j)=\begin{cases}1 & if \quad f^{1,2}\left(v_i^1\right)=v_j^2\\ 0 & otherwise\end{cases} \tag{5}$$

and $\|.\|$ is the Euclidean norm, $\|A\|=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|a_{ij}\right|^2}$.

Note that due to $f^{1,2}$ has to be defined bijective, *P* has only one 1 in each row and column. Thus, the problem of finding the optimal labelling is reduced to the problem of finding the permutation matrix *P* which minimises $C_{f^{1,2}}$. We show in the next section how to find the labelling $f^{1,2}$ using the spectral graph theory.

4.2 Optimal Labelling Between Attributed Graphs

Let $U_{G^1}=\left(u_1^1\middle|u_2^1\middle|\ldots\middle|u_n^1\right)$ and $U_{G^2}=\left(u_1^2\middle|u_2^2\middle|\ldots\middle|u_n^2\right)$ be the modal matrices of the adjacency matrices A_{G^1} and A_{G^2}. Consider that the eigenvectors u_i have been enumerated depending on the value of their eigenvalues, that is, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$.

If we want to find the best labelling $f^{1,2}$ between G^1 and G^2 using the spectral graph theory, we need to project or find a relation between the vertices of the AGs and the spectral decomposition of their adjacency matrices. To do so, we define arbitrarily the bijective functions h^1 and h^2 such that $h^1: v_i^1 \to u_i^1$ and $h^2: v_i^2 \to u_i^2$ and also the diagonal matrices H^1 and H^2 that represent these isomorphisms.

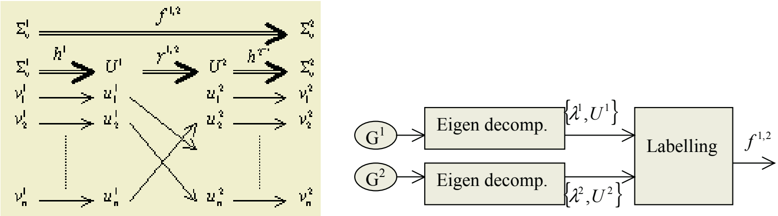


Fig. 1. a) Concatenation of the labelling function between vertices. b) Scheme of the computation of the labelling.

Thus, the function $f^{1,2}$ is defined as the concatenation $f^{1,2}=h^1\circ\gamma^{1,2}\circ h^{2-1}$ where $\gamma^{1,2}$ is a bijective function between the modal matrices U_{G^1} and U_{G^2} (figure 1.a). Moreover, the matrix *P* can be redefined as,

$$P = H^1 M H^{2^T} \tag{6}$$

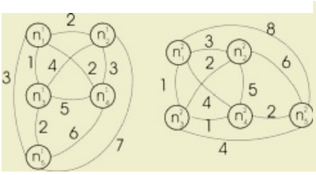
where M represents the isomorphism $\gamma^{1,2}$, that is,

$$M(i,j) = \begin{cases} 1 & \text{if} \quad \gamma^{1,2}(u_i^1) = u_j^2 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

As demonstrated in [6], the global cost $C_{f^{1,2}}$ is minimised when the correlation Γ between the absolute value of the modal matrices U_{G^1} and U_{G^2} is maximised. Then, the permutation matrix M has to be defined such that this correlation is maximised. We can use an algorithm with exponential cost, i.e., the A*, and get the optimal labelling or choose an algorithm with polynomial cost and get a sub-optimal labelling. For instance, applying the Hungarian method to $\overline{U_{G^1}} \cdot \overline{U_{G^2}}^T$ matrix. Figure 1.b shows the basic scheme the method.

4.3 Example of Graph Matching by Spectral Graphs

Assume that we want to compute the distance measure between G^1 and G^2 and decide the optimal labelling between their vertices. Figure 2 shows the AGs G^1 and G^2 and their adjacency matrices. Lines without arrows represent undirected arcs (the attribute in both directions of the edge is the same).



$$A_{G^1} = \begin{pmatrix} 0 & 2 & 1 & 2 & 3 \\ 2 & 0 & 4 & 3 & 7 \\ 1 & 4 & 0 & 5 & 2 \\ 2 & 3 & 5 & 0 & 6 \\ 3 & 7 & 2 & 6 & 0 \end{pmatrix} \qquad A_{G^2} = \begin{pmatrix} 0 & 3 & 1 & 4 & 8 \\ 3 & 0 & 2 & 5 & 6 \\ 1 & 2 & 0 & 1 & 4 \\ 4 & 5 & 1 & 0 & 2 \\ 8 & 6 & 4 & 2 & 0 \end{pmatrix}$$

Fig. 2. The graphs G^1 and G^2 and their adjacency matrices.

The obtained eigenvalues and eigenvectors are

$$\lambda^{G^1} = (14.794, \quad -0.346, \quad -1.974, \quad -3.656, \quad -8.816) \; \lambda^{G^2} = (15.371, \quad 0.036, \quad -1.368, \quad -4.255, \quad -9.783)$$

$$U^{G^1} = \begin{pmatrix} 0.268 & -0.720 & -0.604 & -0.204 & -0.041 \\ 0.496 & -0.064 & 0.499 & -0.494 & -0.505 \\ 0.390 & 0.597 & -0.408 & -0.452 & 0.344 \\ 0.487 & 0.250 & -0.269 & 0.648 & -0.453 \\ 0.539 & -0.239 & 0.380 & 0.297 & 0.646 \end{pmatrix} U^{G^2} = \begin{pmatrix} -0.499 & -0.076 & 0.578 & -0.337 & 0.543 \\ -0.473 & -0.199 & -0.374 & 0.683 & 0.359 \\ -0.264 & 0.742 & -0.489 & -0.328 & 0.177 \\ -0.374 & -0.575 & -0.429 & -0.512 & -0.285 \\ -0.562 & 0.269 & 0.317 & 0.220 & -0.679 \end{pmatrix}$$

Thus, the $\overline{U_{G^1}} \cdot \overline{U_{G^2}}^T$ matrix and the permutation matrix obtained by the Hungarian method is,

$$\overline{U_{G1}} \cdot \overline{U_{G2}} = \begin{pmatrix} 0.6308 & 0.6516 & 0.9761 & 0.8916 & 0.6103 \\ 0.9838 & 0.9542 & 0.6766 & 0.8353 & 0.9077 \\ 0.8174 & 0.8900 & 0.9567 & 0.9959 & 0.8442 \\ 0.8847 & 0.9881 & 0.7411 & 0.9045 & 0.8788 \\ 0.9603 & 0.8813 & 0.7200 & 0.8405 & 0.9938 \end{pmatrix} P = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The labelling that the matrix P represents is $f^{12}(1)=3$, $f^{12}(2)=1$, $f^{12}(3)=4$, $f^{12}(4)=2$ and $f^{12}(5)=5$ and the value of the global cost (3) is: 6.

5 Median Spectral Graphs

Let *Y* be the set of all AGs that can be constructed using the domains of the attribute vertices and arcs (Δ_v, Δ_e). Given a set of AGs $S=\{G^1,G^2,...,G^z\}$, the generalised median graph $\overline{G}$ and the set median graph $\hat{G}$ of *S* are defined in [11] by

$$\overline{G} = \arg\min_{G\in Y} \sum_{i=1}^{z} d\left(G,G^i\right) \qquad \text{and} \qquad \hat{G} = \arg\min_{G\in S} \sum_{i=1}^{z} d\left(G,G^i\right) \tag{8}$$

Both the generalised median graph and the set median graph minimise the sum of distances to all input graphs and the only difference lies in the graph space where the median is searched for. The generalised median is the more general concept and, therefore, usually a better representation of the given patterns than the set median. Notice that $\overline{G}$ is usually not a member of *S*.

We extend the median graph concepts to the spectral theory. Let *K* be the set of all modal matrices. Given a set of modal matrices $L=\{U^1,U^2,...,U^z\}$, we define the generalised median eigenmode $\overline{U}$ and the set median eigenmode $\hat{U}$ of L by

$$\overline{U} = \arg\max_{U\in K} \sum_{i=1}^{z} \Gamma\left(U,U^i\right) \qquad \text{and} \qquad \hat{U} = \arg\max_{U\in L} \sum_{i=1}^{z} \Gamma\left(U,U^i\right) \tag{9}$$

The generalised median eigenmode and the set median eigenmode maximise the sum of the correlations to all modal-matrices in *K* or *L*. Nevertheless, the computation of both medians is drastically different. While $\hat{U}$ is obtained in polynomial time respect the number of elements, $\overline{U}$ is obtained in exponential time.

6 Synthesis of the Generalised Median Eigenmode

Given a set of AGs $S=\{G^1,G^2,...,G^z\}$, which are initially supposed to belong to the same class, we do not, in general, have any way of synthesising the generalised median eigenmode $\overline{U}$ that represents the ensemble unless we can first establish a common labelling of their vertices. We would like to choose the common labelling so as to minimise the measures of dissimilarity between the given AGs and so to maximise the correlation between the median eigenmode $\overline{U}$ and the eigenmodes extracted from the adjacency matrices $L=\{U^1,U^2,...,U^z\}$. This global optimisation problem does not lead to a computationally practical method for choosing the labelling, because there are too many possible orientations to consider, especially when the number and order of the AGs is high.

Two different methods of synthesising a median from a set of elements are used in the literature [2]; the incremental method and the hierarchical method. Therefore, two sub-optimal methods to synthesise $\overline{U}$ could be defined. In the former, $\overline{U}$ is updated by the AGs, which are sequentially introduced. The advantage of this method is that the learning and recognition processes can be interleaved, i.e. the recognition does not need to wait for all the input instances, but is available after each AG has been processed. The main drawback of this incremental approach is that different median eigenmodes can be synthesised from a set of unlabelled AGs depending on the order of presentation of the AGs. To infer some unique $\overline{U}$, a hierarchical method can be defined, which is carried out by successively merging pairs of AGs with minimal distance. The drawback here is that the full ensemble of AGs is needed to generate the median eigenmode. In the experiment results, the median spectral graphs have been obtained by the incremental synthesis.

Algorithm 1 computes the median spectral graph of a set of AGs using an incremental method. It uses three main procedures. The first is the *Eigen_Decomposition* that obtains the eigendecomposition of the adjacency matrix of an AG (section 3). The second is the matching algorithm *M* that finds a sub-optimal labelling between spectral graphs (section 4.2). An the third procedure, *Median_Graph*, updates the median graph. We have to mention that the median function is not transitive, for this reason, the procedure has to keep the information of an AG that is the addition of all the AGs used to compute the median.

```
Algorithm 1: Incremental-synthesis-of-AGs
Inputs: A sequence of AGs $S=\{G^1,G^2,...,G^z\}$, over a common domain.
Output: The eigendecomposition that represents the median graph of $S$.
Begin
    $\{\overline{U},\overline{\lambda}\}$:= Eigen_Decomposition( $G^1$ ) {$\overline{G}$ is composed by only $G^1$}
  for $k := 2$ to $z$ do
      let $f^k : G^k \rightarrow \overline{G}$ be the labelling found by $M\left(U^k,\overline{U}\right)$
      $\overline{G}$ := Median_Graph ( $\overline{G}$ ,$G^k$,$f^k$)
      $\{\overline{U},\overline{\lambda}\}$:= Eigen_Decomposition ( $\overline{G}$ )
end-algorithm
```

7 Experimental Results

In order to examine the behaviour of the new representation, we performed a number of experiments with randomly generated AGs. We randomly generated 10 initial complete AGs, one for each model. From these AGs, the reference and test sets of 10 AGs each were derived by modifying the attribute value of their arcs applying a gausian noise. Figure 3 left shows the ratio of recognition when the noise is increased (Standard deviation from 1 to 14) with 3 different methods: a) *eig2eig*: Comparing graphs using their eigenvalues (section 4). b) *eig2clus*: Comparing graphs to median graphs using eigenvalues (section 5). Approximate median graphs were obtained by the algorithm sketched in section 6. c): *arg2arg*: Comparing graphs using the edit operation matching algorithm described in [13]. Only the substitution edit operation has been considered since we want to find the best bijective labelling between vertices of both graphs.

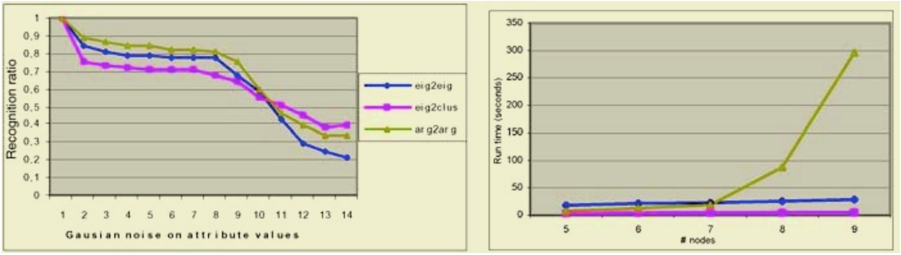


Fig. 3. Recognition ratio and run time in seconds of randomly generated attributed graphs.

It is interesting to emphasize that the ratio of recognition of the spectral methods is only slightly lower than the edit operation methods although the cost of the first ones are polynomic and the second ones are exponential. Considering the spectral methods, we have to stress that in the *eig2clus* method, the run time (figure 3 right) of the recognition depends only on the number of clusters but not on the number of AGs per cluster. For this reason, in these experiments, the run time of the *eig2clus* method were 10 times lower than the *eig2eig* method. Moreover, when the AGs in a cluster are very different (the noise is high), the *eig2clus* method keeps the structural information of the cluster and for this reason the recognition ratio obtains the best results.

8 Conclusions and Future Work

In this paper, we have merged the median computation with the spectral graph theories to define the Median Spectral Graphs. The aim of this new structure is to obtain the advantage of the structural pattern recognition but with a polynomic computational cost. By decomposing the adjacency matrices of the graphs in their eigenvectors and eigenvalues, we obtain a suboptimal labelling in polynomic cost. Experimental results show that this scheme is useful to keep the structural information of the graphs of each class when they are synthesised in only one median structure. As a future work, we have to define the hierarchical synthesis and test our new methods in a real application.

References

1. A.K.C. Wong & M. You, Entropy and distance of random graphs with application to structural pattern recognition, IEEE Trans. on PAMI, vol. 7, pp. 599-609, 1985.
2. F. Serratosa, R. Alquézar & A. Sanfeliu, Synthesis of Function-Described Graphs and clustering of Attributed Graphs, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 16, No. 6, pp. 621-655, 2002.
3. A. Sanfeliu, F. Serratosa & R. Alquézar, Second-Order Random Graphs for modeling sets of Attributed Graphs and their application to object learning and recognition, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 18, No. 3, pp: 375-396, 2004.
4. H. Bunke and B. Messmer, Recent advances in graph matching. International Journal in Pattern Recognition and Artificial Intelligence, vol. 11, pp. 169-203, 1997.
5. B. Mohar, Laplace Eigenvalues of Graphs – A survey, Discrete Mathematics 109, pp: 171-183, 1992.
6. S. Umeyama, An eigen decomposition approach to weighted graph matching problems, IEEE Trans. on Pattern Analysis and Machine Intelligence 10, pp: 695-703, 1988.
7. L.Xu and I. King, A PCA Approach for Fast Retrieval of Structural Patters in Attibuited Graphs, IEEE Trans. Systems, Man and Cybernetics, vol. 31. NO. 5, pp 812-817, 2001.
8. L. Xu, and A. Yuille, Robust principal component analysis by self-organizing rules based on statistical physics approach, IEEE Trans. Neural Networks, vol. 6, pp. 131-143, 1995.
9. R.C. Wilson and E.R. Hancock, Spectral Analysis of Complex Laplacian Matrices, LNCS 3138, pp:57-65, 2004.
10. B. Luo, R.C. Wilson and E.R. Hancock, Spectral embedding of graphs, Pattern Recognition 36, pp:2213-2230, 2003.
11. X. Jiang, A. Münger & H. Bunke, On Median Graphs: Properties, Algorithms and Applications, Transactions on Pattern Analysis and Artificial Intelligence, Vol. 23, No. 10, pp. 1144-1151, 2001.
12. J.Weng, Y.Zhang & W-S. Hwang, Candid Covariance-Free Incremental Principal Component Analysis, IEEE Trans. on PAMI, vol. 25, No. 8, pp. 1034-1040, 2003.
13. A. Sanfeliu & K. Fu, A distance measure between attributed relational graphs for pattern recognition, IEEE Transactions on Systems, Man and Cybernetics, vol. 13, pp. 353-362, 1983.

Feature Selection for Graph-Based Image Classifiers

Bertrand Le Saux and Horst Bunke

Institut für Informatik und Angewandte Mathematik,
University of Bern, Neubrückstrasse, 10, CH-3012, Bern, Switzerland
{lesaux,bunke}@iam.unibe.ch

Abstract. The interpretation of natural scenes, generally so obvious and effortless for humans, still remains a challenge in computer vision. We propose in this article to design binary classifiers capable to recognise some generic image categories. Images are represented by graphs of regions and we define a graph edit distance to measure the dissimilarity between them. Furthermore a feature selection step is used to pick in the image the most meaningful regions for a given category and thus have a compact and appropriate graph representation.

1 Introduction

How can one construct computer programmes in order to understand the content of scenes? Such programmes would satisfy needs in image retrieval and computer vision, and could possibly be applied to a wide range of areas, including security, digital libraries and web searching. We propose in this article to design binary classifiers capable to recognise some generic image categories.

Previously, image classification has been performed by using directly support vector machines on image histograms [1] or hidden Markov models on multiresolution features [2]. These methods do not take into account that human description of an image content is rarely global but often specific to an image part. To include local information, attributed relational graphs [3] and image blocks [4] were proposed. Such approaches rely on the ability of the classifiers to distinguish between complex features, so they are prone to over-fit when the concept to learn has a large variance.

Our approach segments images into regions and index each image by a graph of regions. For a given type of scene, only image parts that are meaningful in that case are selected in order to make easier the task of the classifiers. This allows to define an efficient comparison scheme between the graphs that represent the images.

This paper is organised as follows. In Sect. 2, we explain how to describe the images and how to select the meaningful regions. The graph matching procedure and the classification process are described in Sect. 3. Finally, we present some experiments and discuss their results in Sect. 4.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 147–154, 2005.

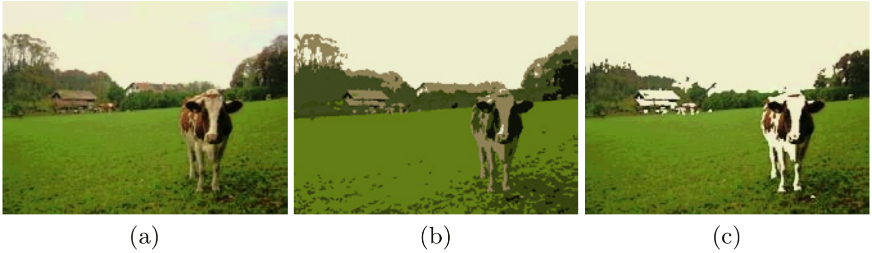


Fig. 1. The original image (a) is first segmented (b), and then we keep only the regions (c) the type of which has a large mutual information with the scene to predict. When indexing this image with respect to the *countryside* label, the sky and the buildings are considered as non-informative and are discarded.

2 Image Representation

2.1 From Images to Regions

Images are first segmented into regions by using the mean shift algorithm [5]. This is a simple non-parametric technique for maximisation of the probability density. It basically performs a density gradient ascent.

To perform colour segmentation, the mean shift procedure is applied at various start locations, then the obtained high density colours are mapped to the image plane to keep only those belonging to large enough regions. Typically, this technique gives results as shown on Fig. 1: there are less than 10 regions per image, that are not necessarily connected but correspond more or less to the main semantic areas since colour is an important visual cue for generic images.

2.2 Feature Selection

Region Lexicon. The region lexicon consists of a list of the region types that occur in an image data set. Such a data set is built by gathering various generic images. Once they are segmented, these images are assumed to provide a good representation of the possible image regions that occur in the real world.

We cluster this data set of image regions using techniques previously proposed to find clusters of visually similar images in image databases [6] and based on fuzzy clustering methods [7]. The resulting clusters contain visually similar image regions and thus define implicitly a region type. Each of them is included in the region lexicon.

Selection of Meaningful Regions. For a segmented image, we can determine the type of each region simply by computing the distance (based on the region descriptor) to the cluster centroids and choosing the closest one. Let $\mathcal{I}$ denote the set of images, and X a random variable on $\mathcal{I}$ standing for the distribution of images. We can build a set of features $F = \{f_1, \dots, f_N\}$ which are mappings from $\mathcal{I} \to \{0, 1\}$. In the experiments those features are indicators of the presence - or absence - of a given region type in the image. We denote $F_1 = f_1(X), \dots, F_p = f_p(X)$ the boolean random variables associated with those features.

In order to understand which region types are meaningful to recognise a concept, a filtering phase based on feature selection [8] is applied as in [9]. The most standard ways to select features consist in ranking them according to their individual predictive power, that may be estimated by mutual information [10].

Information theory [11] provides tools to assess the available features. The entropy measures the average number of bits required to encode the value of a random variable. For instance, if we denote Y a boolean random variable standing for the class to predict (i.e. the concept to associate with the image), its entropy is $H(Y) = -\sum_y P(Y = y)\log(P(Y = y))$. The conditional entropy $H(Y|F_j) = H(Y, F_j) - H(F_j)$ quantifies the number of bits required to describe Y when the feature F_j is already known. The mutual information of the class and the feature quantifies how much information is shared between them and is defined by:

$$\begin{aligned} I(Y, F_j) &= H(Y) - H(Y|F_j) \\ &= H(Y) + H(F_j) - H(Y, F_j) \end{aligned} \tag{1}$$

The features f_j are ranked according to the information $I(Y, F_j)$ they convey about the class to predict, and those with the largest mutual information are chosen. In the image, we keep only the regions that have a region type among the selected ones (cf. Fig. 1). They are the most meaningful ones to recognise the concept.

2.3 From Regions to Graphs

Definition 1. *A graph G is a 4-tuple $G = (V, E, \mu, \nu)$ where*

- *V is the set of vertices;*
- *$E \subseteq V \times V$ is the set of edges;*
- *$\mu : V \rightarrow L_V$ is a function assigning labels to the vertices;*
- *$\nu : E \rightarrow L_E$ is a function assigning labels to the edges.*

Two different alternatives to represent images by a graph are investigated in this paper. In either case, each region constitutes a vertex of the graph. The vertex labels are the colour histograms that characterise the corresponding region. In the first type of graph representation, only vertices corresponding to adjacent regions (i.e. with at least one point of contact) are linked by an edge, with no label. In the second graph representation, all vertices are linked to all the other ones, with a label defined proportionally to the common boundary length (CBL). Both types of graphs are undirected.

3 Image Classification

Classification of images implies to be able to measure the similarity between the graphs representing the images. Moreover in the case of images, data are usually corrupted by noise and strongly depend on illumination conditions. Error correcting methods for graph matching have been proposed to cope with these problems. Among them, the graph edit distance is particularly popular. It defines a set of possible edit operations and assigns a cost to each of them. The distance of two graphs is then the minimum cost of all sequences of edit operations that transform a graph into the other. To compute the graph edit distance, we use the A^* algorithm [12] as described for graph matching in [13]. A look-ahead procedure [14] is used to speed up the matching process. Last, a k-Nearest-Neighbour (k-NN) classifier is used to classify the images. Next sections describe the edit operations and their associated cost.

3.1 Graph Edit Operations

Let p be a mapping between the vertices of two graphs $G_1 = (V_1, E_1, \mu_1, \nu_1)$ and $G_2 = (V_2, E_2, \mu_2, \nu_2)$. We assume that G_1 and G_2 are such that $\mathrm{Card}(V_1) \leq \mathrm{Card}(V_2)$. This mapping consists of elementary mappings (v, w), $v \in V_1$ and $w \in V_2$ such that each vertex is used only once. The \$ element denotes a missing vertex in graph G_2. For each couple (v, w) in p, the possible vertex edit operations are defined as follows:

- vertex label substitution: if $w \neq \$$ the mapping implies the substitution of $\mu_1(v)$ by $\mu_2(w)$.
- vertex deletion: if $w = \$$, it implies the deletion of v from G_1.

For each pair of elementary mappings (v, w) and (v', w') in p, the possible edge edit operations are defined as follows:

- edge label substitution: if $\exists$ an edge $e_1 = (v, v') \in E_1$ and an edge $e_2 = (w, w') \in E_2$, the mapping implies the substitution of edge label $\nu_1(v, v')$ by $\nu_2(w, w')$.
- edge deletion: if $\exists$ an edge $e_1 = (v, v') \in E_1$ and there is no edge $(w, w') \in E_2$, it implies the deletion of e_1 from E_1.
- edge insertion: if there is no edge $(v, v') \in E_1$ but $\exists$ an edge $e_2 = (w, w') \in E_2$, then $e_1 = (v, v')$ has to be inserted in E_1.

3.2 Graph Edit Costs

Different sets of graph edit costs are defined for the two graph representations of images defined in Sect. 2.3. In both cases, the vertices convey the visual information about the image regions, so the vertex edit operations have the same cost:

Definition 2. *Vertex edit costs for both graph representations:*

- *vertex label substitution: the cost of the substitution of $\mu_1(v)$ by $\mu_2(w)$ is the Euclidean distance between the labels (i.e. the colour histograms of the image regions):* $c(\mu_1(v) \rightarrow \mu_2(w)) = ||\mu_1(v) - \mu_2(w)||_2$.
- *vertex deletion: to make the deletion easier on large graphs than on small ones:* $c(v \rightarrow \$) = \frac{1}{\mathrm{Card}(V_1)}$.

In the first graph representation, graphs have only edges corresponding to adjacent regions and the corresponding costs are defined as:

Definition 3. *Edge edit costs for set #1:*

- *edge label substitution: by definition there is a perfect match so there is no cost:* $c(\nu_1(e_1) \rightarrow \nu_2(e_2)) = 0$.
- *edge deletion: to take into account the size of the graph and have comparable costs:* $c(e_1 \rightarrow \$) = \frac{1}{\mathrm{Card}(V_1)}$.
- *edge insertion: by symmetry:* $c(\$ \rightarrow e_1) = \frac{1}{\mathrm{Card}(V_1)}$.

A second way to define the edges is based on the the common boundary length (CBL) of two regions. The edge label could be defined as the CBL itself, or a normalised value based on the CBL, for example $\max(\frac{CBL}{BL_{\mathrm{reg1}}}, \frac{CBL}{BL_{\mathrm{reg2}}})$ or $\mathrm{avg}(\frac{CBL}{BL_{\mathrm{reg1}}}, \frac{CBL}{BL_{\mathrm{reg2}}})$ where $BL_{\mathrm{reg}i}$ is the boundary length of region i. For such graphs, since there exist edges between all pairs of vertices, there is no need anymore for edge deletion or insertion operations:

Definition 4. *Edge edit costs for set #2.*

- *edge label substitution: for any pair of edges e_1 and e_2,*

$$c(\nu_1(e_1) \rightarrow \nu_2(e_2)) = ||\nu_1(e_1) - \nu_2(e_2)||_2$$

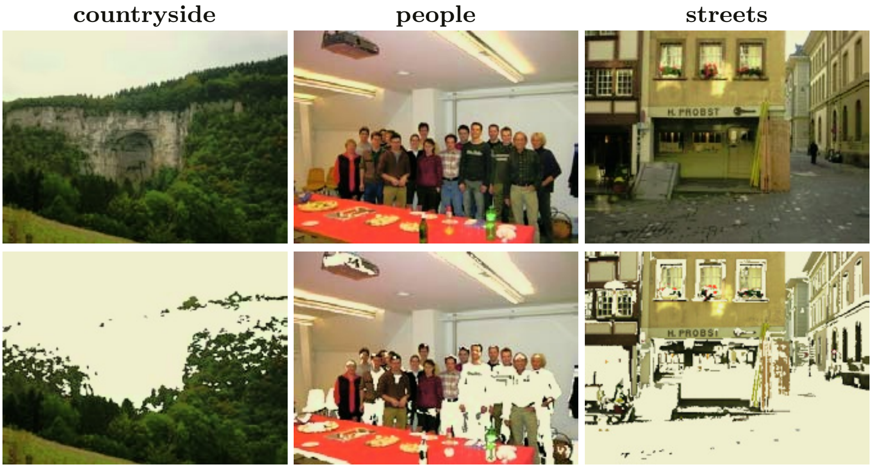
4 Experiments

4.1 Data Set

The data set is composed of 200 images collected from the web. Four classes contain instances of a particular scene type: *snowy, countryside, streets* and *people*. A fifth one consists of various generic images aimed to catch a glimpse of the possible real scenes and thus used as negative samples for the classifiers. In the experiments, training categories of 30 instances are extracted randomly from the data set and error rates are averaged on 25 runs. Some examples are shown in Fig. 2.

4.2 Graph Matching Classification

The edit cost sets proposed in Sect. 3.2 are compared in Table 1. The quality of the considered set of edit costs depends on the complexity of the underlying scenes. For class *snowy* which is rather easy to recognise, the second set of edit costs is superior to the first one. However, for class *people* which is rather difficult, the situation is just the opposite. Since we intend to build some generic classifiers able to recognise different types of scenes, they have to satisfy an overall criterion including both the smallest average error and the smallest standard deviation. The last graph representation has the best test error rate on average, but shows



countryside **people** **streets**

Fig. 2. The meaningful regions correspond to the region types that have a high mutual information with the label to predict. The upper row shows the original images and the lower row shows only the meaningful regions in these images.

Table 1. Error rates for various keywords: comparison of various edit cost sets.

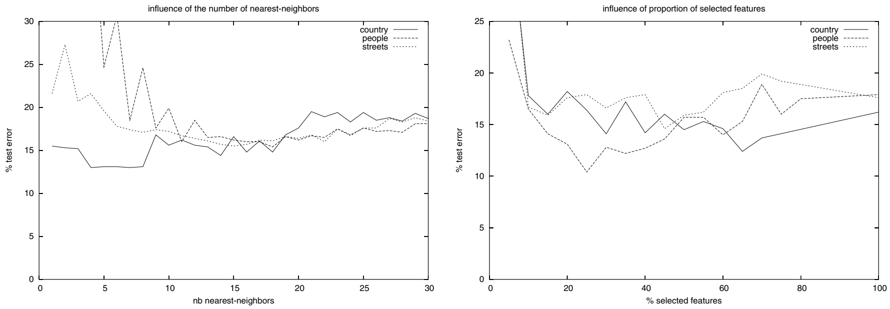
keyword	edit cost set #1		edit cost set #2 ($\nu = CBL$)		edit cost set #2 ($\nu = max(\frac{CBL}{BL_1}, \frac{CBL}{BL_2})$)	
	training error	test error	training error	test error	training error	test error
snowy	8.4 %	11.4 %	9.4 %	8.2 %	9.1 %	7.9 %
country	14.5 %	16.3 %	16.5 %	15.8 %	15.4 %	14.4 %
people	12.8 %	15.6 %	19.1 %	20.9 %	17.5 %	19.4 %
streets	14.6 %	17.3 %	16.9 %	16.0 %	17.3 %	15.3 %
mean		15.15 %		15.22 %		13.75 %
deviation		2.25 %		4.54 %		4.15 %

large disparities between scenes. We observe that edge labels based on the simple adjacency between the regions result in the best overall performance.

Figure 3-a illustrates the influence of the number of neighbours in the classifier on the test error rate. For the complex scenes, the graphs indicate there exists an optimal value (around 15 neighbours). This is less obvious on simpler scenes like *country*, for which error rates are rather constant, with a slight trend to increase with the number of neighbours. The number of neighbours is then set to 15 for the complex scenes and 5 for the simple ones.

4.3 Influence of the Feature Selection

For each type of scene, the feature selection allows to pick out the meaningful parts of the image. Figure 2 shows the selected regions are consistent with what can be expected intuitively. The influence of the proportion of selected features



(a) (b)

Fig. 3. (a) The number of neighbours has more influence on the complex scenes for which an optimal value can be found with roughly 15 neighbours. (b) A feature selection rate between a third and a half of the features allows to obtain the best error rates.

on the test error rate is presented in Fig. 3-b. The graphs show that lower error rates can be obtained by selecting roughly between a third and a half of the region types: the optimal values are then chosen as a tuning reference for each concept. Table 2 compares the error rates with and without region selection: performance is improved for each category.

Table 2. Influence of the feature selection on the error rates.

keyword	with all regions		with region selection	
	training error	test error	training error	test error
snowy	8.4 %	11.4 %	11.1 %	10.9 %
country	14.5 %	16.3 %	14.5 %	12.4 %
people	12.8 %	15.6 %	12.7 %	10.4 %
streets	14.6 %	17.3 %	17.1 %	14.6 %

Table 3 compiles the computing times of the graph matching process and the image classification task (performed on a computer with a 800 MHz processor) for various proportions of selected features. Since the algorithm complexity is exponential with the number of vertices, feature selection appears as an intelligent way to greatly speed up the process.

Table 3. Influence of the feature selection on the computational costs.

	nodes	graph distance	classif.
all features	5.6	74.8 ms	7659.1 ms
66% features	3.9	5.1 ms	497.0 ms
50% features	3.0	1.6 ms	117.8 ms
33% features	2.3	0.3 ms	47.6 ms

5 Conclusion

In this article we have presented a new approach for image classification, which is based on a graph representation of the images. The classifier is a k-Nearest-Neighbour algorithm and uses a graph edit distance for which we have evaluated different sets of edit costs to find the most appropriate one for image analysis.

Furthermore, we have shown that a region selection by maximisation of the mutual information between the region types and the class to predict greatly improves the recognition rates while reducing the complexity of the graph matching. This allows the classifier to offer competitive computing times.

Other existing methods in the literature stress different features in the image. For instance [1] or [9] lead to more or less comparable results, but what is more, our method performs better on the type of scenes that are difficult for them. Further work will investigate how our approach can be combined with these ones to achieve a better overall performance.

References

1. Chapelle, O., Haffner, P., Vapnik, V.: SVMs for histogram-based image classification. IEEE Transactions on Neural Networks **10** (1999) 1055–1065
2. Li, J., Wang, J.Z.: Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. PAMI **25** (2003) 1075–1088
3. Beretti, S., Del Bimbo, A., Vicario, E.: Efficient matching and indexing of graph models in content-based retrieval. IEEE Trans. PAMI **23** (2001) 1089–1105
4. Minka, T., Picard, R.: Interactive learning using a society of models. Pattern Recognition **30** (1997) 565–581
5. Comaniciu, D., Meer, P.: Robust analysis of feature spaces: Color image segmentation. In: Proceedings of CVPR, San Juan, Porto Rico (1997) 750–755
6. Le Saux, B., Boujemaa, N.: Unsupervised robust clustering for image database categorization. In: Proceedings of ICPR, Quebec, Canada (2002) 259–262
7. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New-York, N.Y. (1981)
8. Guyon, I., Elisseff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research **3** (2003) 1157–1182
9. Le Saux, B., Amato, G.: Image recognition for digital libraries. In: ACM Multimedia/International Workshop on Multimedia Information Retrieval. (2004)
10. Battiti, R.: Using mutual information for selecting features in supervised neural network learning. Neural Networks **5** (1994) 537–550
11. Gray, R.M.: Entropy and Information Theory. Springer-Verlag, New York, N.Y. (1990)
12. Nilsson, N.J.: Principles of Artificial Intelligence. Tioga, Palo Alto, CA (1980)
13. Messmer, B.: Graph Matching Algorithms and Applications. PhD thesis, University of Bern (1995)
14. Wong, E.: Three-dimensional object recognition by attributed graphs. In Bunke, H., Sanfeliu, A., eds.: Synctatic and Structural Pattern Recognition - Theory and Applications. (1990) 381–314

Machine Learning with Seriated Graphs

Hang Yu and Edwin R. Hancock

Department of Computer Science, University of York, York Y01 5DD, UK

Abstract. The aim in this paper is to show how the problem of learning the class-structure and modes of structural variation in sets of graphs can be solved by converting the graphs to strings. We commence by showing how the problem of converting graphs to strings, or seriation, can be solved using semi-definite programming (SDP). This is a convex optimisation procedure that has recently found widespread use in computer vision for problems including image segmentation and relaxation labelling. We detail the representation needed to cast the graph-seriation problem in a matrix setting so that it can be solved using SDP. We show how the strings delivered by our method can be used for graph-clustering and the construction of graph eigenspaces.

1 Introduction

The problem of placing the nodes of a graph in a serial order is an important practical problem that has proved to be theoretically difficult. The task is one of practical importance since it is central to problems such as network routing, the analysis of protein structure and the visualisation or drawing of graphs. Moreover, and of central importance to this paper, if the nodes of graphs can be placed in a serial order then conventional machine learning methods may be applied to them. Theoretically, the problem is a challenging one since the problem of locating optimal paths on graphs is one that is thought to be NP-hard [8]. The problem is known under a number of different names including "the minimum linear arrangement problem" (MLA) [10] and "graph-seriation"[5].

Stated formally, the problem is that of finding a permutation of the nodes of a graph that satisfies constraints provided by the edges of the graph. The recovery of the permutation order can be posed as an optimisation problem. It has been shown that when the cost-function is harmonic, then an approximate solution is given by the Fiedler vector of the Laplacian matrix for the graph under study [5]. Thus, the solution to the seriation problem is closely akin to that of finding a steady state random walk on the graph, since this too is determined by the Laplacian spectrum. However, the harmonic function does not necessarily guarantee that the nodes are arranged in an order that maximally preserves edge connectivity constraints. In a recent paper, Robles-Kelly and Hancock [1] have reformulated the problem as that of recovering the node permutation order subject to edge connectivity constraints, and have provided an approximate spectral solution to the problem.

Although spectral methods are elegant and convenient, they are only guaranteed to locate solutions that are locally optimal. Recently, semidefinite programming (SDP) [7] has been developed as an alternative method for locating optimal solutions couched

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 155–162, 2005.

in terms of a matrix representation. Broadly speaking, the advantage of the method is that it has improved convexity properties, and is less likely to locate a local optimum. The method has been applied to a number of graph-based problems in pattern recognition including graph partitioning [3], segmentation [4] and the subgraph isomorphism problem [2].

The aim in this paper is hence to investigate whether SDP can be applied to the graph-seriation problem and whether the resulting strings can be used for machine learning. We commence by illustrating how the cost-function can be encoded in a matrix form to which SDP can be applied. With this representation to hand, then standard SDP methods can be applied to extract the optimal serial ordering. To do this we lift the cost function to a higher-dimensional space. Here the optimization problem is relaxed to one of convex optimization, and the solution recovered by using a small set of random hyperplanes. We explore how the resulting strings delivered by the seriation method can be used for the purposes of learning the class structure (i.e. graph clustering) and determining the modes of structural variation present for graphs of a particular class.

2 Graph Seriation

We are concerned with the undirected graph $G = (V, E)$ with node index-set V and edge-set $E =\subseteq V \times V$. The adjacency matrix A for the graph is the $V \times V$ matrix with elements

$$A(i,j) = \begin{cases} 1 & \text{if}(i,j) \in E \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The graph seriation problem has been formally posed as one of optimisation in the work of Atkins *et al* [5]. Formally, the problem can be stated as finding a path sequence for the nodes in the graph using a permutation π which will minimize the penalty function

$$g(\pi) = \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} A(i,j)(\pi(i) - \pi(j))^2 \tag{2}$$

Since the task of minimizing g is NP-hard due to the discrete nature of the permutation, a relaxed solution is sought using a function h of continuous variables x_i. The relaxed problem can be posed as seeking the solution of the constrained optmisation problem $x = \arg\min_{x^*} h(x^*)$ where $h(x) = \sum_{(i,j)} f(i,j)(x_i - x_j)^2$ subject to the constraints $\sum_i x_i = 0$ and $\sum_i x_i^2 = 1$. Using graph-spectral methods, Atkins and his coworkers showed that the solution to the above problem can be obtained from the Laplacian matrix of the graph. The Laplacian matrix is defined to be $L_A = D_A - A$ where D_A is a diagonal matrix with $d_{i,i} = \sum_{j=1}^{n} A_{i,j}$. The solution to the relaxed seriation problem is given by the Fiedler vector, i.e. the vector associated with the smallest non-zero eigenvalue of L_A. The required serial ordering is found by sorting the elements of the Fiedler vector into rank-order. Recently, Robles-Kelly and Hancock [1] have extended the graph seriation problem by adding edge connectivity constraints. The graph seriation problem is restated as that of minimising the cost-function

$$h_E(x) = \sum_{i=1}^{|V|-1} \sum_{k=1}^{|V|} (A(i,k) + A(i+1,k))x_k^2 \tag{3}$$

By introducing the matrix

$$\Omega = \begin{bmatrix} 1 & 0 & 0 & 0 \cdots 0 & 0 \\ 0 & 2 & 0 & 0 \cdots 0 & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & 0 \cdots 2 & 0 \\ 0 & 0 & 0 & 0 \cdots 0 & 1 \end{bmatrix}$$

the path connectivity requirement is made more explicit. The minimiser of $h_E(\boldsymbol{x})$ satisfies the condition

$$\lambda = arg \min_{x_*} \frac{\boldsymbol{x_*}^T \Omega A \boldsymbol{x_*}}{\boldsymbol{x_*}^T \Omega \boldsymbol{x_*}} \tag{4}$$

Although elegant and convenient, spectral methods are only guaranteed to find a locally optimal solution to the problem. For this reason in this paper we turn to the more general method of semidefinite programming to locate an optimal solution which utilizes the convexity properties of the matrix representation.

3 Semidefinite Programming

Semidefinite programming (SDP) is an area of intense current topical interest in optimization. Generally speaking, the technique is one of convex optimisation that is efficient since it uses interior-point methods. The method has been applied to a variety of optimisation tasks in combinatorial optimization, matrix completion and dual Lagrangian relaxation on quadratic models. Semidefinite programming is essentially an extension of ordinary linear programming, where the vector variables are replaced by matrix variables and the nonnegativity elementwise constraints are replaced by positive semidefiniteness. The standard form for the primal problem is: $X = \arg\min_{X^*}$ $traceCX^*$, such that $traceF_iX = b_i$, $i = 1...m$, $X \succeq 0$. Here C, F_i and X are real symmetric $n \times n$ matrices and b_i is a scalar. The constraint $X \succeq 0$ means that the variable matrix must lie on the closed convex cone of positive semidefinite solutions. To solve the graph seriation problem using semidefinite programming, we denote the quantity $\Omega^{1/2} A \Omega^{-1/2}$ appearing in equation (4) by B and $\Omega^{1/2} x_*$ by y. With this notation the optimisation problem can be restated as $\lambda = \arg\min_{y^T y=1} y^T By$. Noting that $y^T By = trace(Byy^T)$ by letting $Y = yy^T$ in the semidefinite programming setting the seriation problem becomes $Y = \arg\min_{Y^*} traceBY^*$ such that $traceEY^* = 1$, where the matrix E is the unit matrix, with the diagonal elements set to 1 and all the off-diagonal set to 0. Note that $Y = yy^T$ is positive semidefinite and has rank one. As a result it is convex and we can add the positive semidefinite condition $Y \in S_n^+$ where S_n^+ denotes the set of symmetric $n \times n$ matrices which are positive semidefinite.

3.1 Interior Point Algorithm

To compute the optimal solution Y^*, a variety of iterative interior point methods can be used. By using the SDP solver developed by Fujisawa et.al [6], a primal solution matrix Y^* can be obtained. Using the solution Y^* to the convex optimization problem (**??**), we must find an ordered solution y to the original problem. To do this we use the randomized-hyperplane technique proposed by Goemans and Williamson [9].

Since $Y^* \in S_n^+$, by using the Cholesky decomposition we have that $Y = V^T V, V = (v_1, ...v_n)$.Recalling the constraint $y^T y = 1$, the vector y must lie on the unit sphere in a high dimensional space. This means that we can use the randomized hyperplanes approximation. This involves choosing a random vector r from the unit sphere. An ordered solution can then be calculated from $Y^* = V^T V$ by ordering the value of $v_i^T r$. We repeat this procedure multiple times for different random vectors. The final solution y_* is the one that yields the minimum value for the objective function $y^T By$. This technique can be interpreted as selecting different hyperplanes through the origin, identified by their normal r, which partition the vectors $v_i, i = 1....n$.

The solution vector x_* can be obtained using the equation $\Omega^{1/2} x_* = y$, and the elements of the vector x_* then can be used to construct the serial ordering of the nodes in the graph. Commencing from the node associated with the largest component of x_*, we sort the nodes in so that the nodes are ordered so that the components of x_* are of decreasing magnitude and also satisfy edge connectivity constraints on the graph. We iteratively proceed in the following. Let us denote the list of the visited nodes by S_k at the kth iteration. Initially $S_1 = i_1 = \arg\max_i x_*(i)$. We proceed by searching the set of the first neighbours of i_1, i.e. $N_{i_1} = \{j | (i_1, j) \in E\}$, to locate the node which is associated with the largest remaining component of x_*. This node is then appended to the list of nodes visited list and satisfies the condition $i_2 = \arg\max_{l \in N_{i_1}} x_*(l)$. This process is repeated until every node in the graph is visited. At termination the sorted list of nodes is the string S_G.

4 Graph Matching

With the converted strings at hand, we are able to pose the graph matching problem as that of aligning the strings so as to minimise the transition cost on a string edit matrix. We denote the seriations of the data graph $G_D = (V_D, E_D)$and model graph $G_M = (V_M, E_M)$ by $X = \{x_1, x_2, ..., x_m\}$ and $Y = \{y_1, y_2, ..., y_n\}$ respectively. Here m and n represent the number of nodes in the two graphs. These two strings can be used to index the rows and columns of an edit lattice. Since the graphs may have different sizes, we introduce a null symbol ϵ which can be used to pad the strings. The graph matching problem can be stated as finding a path $\Gamma =< p_1, p_2, ...p_k..., p_L >$ through the lattice which generates the minimum transition cost. Each element $p_k \in (V_D \cup \epsilon) \times (V_M \cup \epsilon)$ of the edit path is a Cartesian pair. We constrain the path to be connected on the edit lattice, and also the transition from the state p_k to the state p_{k+1} is constrained to move in a direction on the lattice, which is increasing and connected in the horizontal, vertical or diagonal directions on the lattice. The diagonal transition corresponds to the match of an edge of the data graph to an edge of the model graph. A horizontal transition implies that the traversed nodes of the model graph are null-matched. Similarly, the visited nodes of the data graph are null-matched when a vertical transition is made.

By representing the adjacent states on the path by p_k and p_{k+1}, the cost function of the edit path can be given as follows:

$$d(X,Y) = \sum_{p_k \in \Gamma} \eta(p_k \rightarrow p_{k+1}) \tag{5}$$

where $\eta(p_k \rightarrow p_{k+1})$ is the transition cost between the adjacent states. The optimal edit path is the one that minimises the edit distance between string and satisfies the condition $\Gamma^* = arg \min_{\Gamma} d(X,Y)$. The optimal edit sequence may be found using Dijkstra's algorithm and the matching results are obtained from the optimal transition path on the edit lattice.

5 Computing a Reference String

We are interested in whether the strings delivered by our graph seriation method can be used for the purposes of graph clustering and constructing eigenspaces for graphs. To do this a reference string is required, since this can be used as a class prototype, and also allows the covariance matrix for a set of strings (i.e. seriated graphs) to be computed. To construct the reference string, we proceed as follows. After converting the set of M graphs $\{G_1, G_2, .., G_k, ..G_M\}$ into a set of strings $\{S_{G_1}, S_{G_2}, .., S_{G_k}, .., S_{G_M}\}$, we compute the pair-wise edit distances of the strings using the correspondences between graphs obtained using graph matching technique. We denote the edit distance matrix by ED_G. We then select the reference string $S_{\{r\}}$ so as to satisfy the condition $r = \arg\min_{r^*} \sum_{j \in |M|} ED_G(r^*, j)$.

This reference string can be used to capture the statistical properties of the set of graphs. In order to create a meaningful pattern-space for graph clustering, we construct permuted graph adjacency matrices by making use of the matching results between the individual string $S_{G-\{r\}}$ and the reference string S_r. For the graph indexed k, the permuted adjacency matrix is given by

$$\mathcal{A}_k(i,j) = \begin{cases} 1 & \text{if } (C(i), C(j)) \in E \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where the $C(i)$ and $C(j)$ represent the node correspondences of nodes i and j in the reference string. Next we convert the permuted adjacency matrices into long-vectors by stacking the columns of the permuted adjacency matrices. For the graph indexed k, the long vector is $H_k = (\mathcal{A}_k(1,1), \mathcal{A}_k(2,1), \mathcal{A}_k(3,1), ...)^T$.

Our aim is to construct an eigenspace which can be used to capture the modes of variations is graph edge-structure. To do this, we represent the variations present in the set of graphs using the mean long-vector and the covariance matrix for the long-vectors. The eigenspace is constructed by projecting the individual graph long-vectors onto the directions spanned by the principal eigenvectors of the covariance matrix.

To be more formal, we commence by calculating the mean long-vector ($\hat{z}$) and the long-vector covariance matrix (σ) for the set of permuted adjacency matrices using the following formulae

$$\hat{H} = \frac{1}{M}\sum_{k=1}^{M} H_k \qquad \Sigma = \frac{1}{M}\sum_{k=1}^{M}(H_k - \hat{H})(H_k - \hat{H})^T. \tag{7}$$

To construct the eigenspace we commence by computing the eigenvalues and eigenvectors for the covariance matrix Σ. The eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$ are found by solving the polynomial equations $|\Sigma - \lambda I| = 0$, where I is the identity matrix. The associated eigenvectors $\phi_1, \phi_2, \ldots, \phi_N$ are found by solving the linear eigenvector equation $\Sigma\phi_k = \lambda_k\phi_k$. From the eigenvectors we construct a modal matrix. The eigenvectors are ordered in decreasing eigenvalue order to form the columns of the modal matrix, denoted by $\Phi = (\phi_1|\phi_2|\ldots|\phi_N)$. If eigenspace is taken over the leading K eigenvectors, then the projection matrix is $\Phi_K = (\phi_1|\phi_2|\ldots|\phi_M)$. The projection of the long-vector H_k onto the eigenspace is given by $\mathcal{H}_k = \Phi_K^T H_k$.

6 Experiments

In this section, we provide an experimental evaluation of our new algorithm for graph seriation. Our experimental evaluation is divided into two parts. First, we present results for the clustering of graphs using edit distances between seriated node sequences. In the second part, we test the utility of reference string as a class prototype. For our experimental evaluation we use the COIL image database. To extract graphs from the images, we first detect feature points using the Harris corner detector. The graphs used in our study are the Delaunay triangulations of the point sets. The reason for using Delaunay graph is that it incorporates important structural information from the original image. In the images studied there are rotation, scaling and perspective distortions present. Example images from the sequences are shown in Fig 1 and correspond to different camera viewing directions of the objects. The detected feature points and their Delaunay triangulations are overlayed on the images.

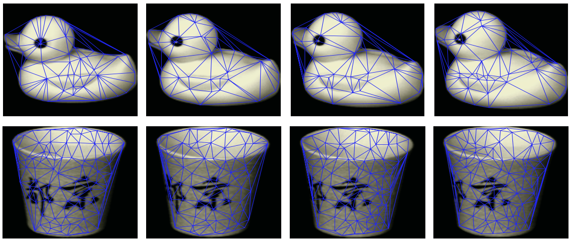


Fig. 1. Delaunay graphs overlayed on coil data.

To explore the clustering of graphs, we have selected four objects from the COIL database. For each object there are 72 different views. For the 288 graphs in the dataset, we have computed the complete set of distances between each pair of graphs. We have performed clustering the graphs using the following procedure. First, we convert the graphs into strings using our SDP seriation method. Second, the pair-wise correspondences between two different graphs in the set are located. Finally we compute the edit distances by using the correspondences on the serialized strings. With the edit distance matrix at hand, we apply the multidimensional scaling to the edit distance matrix. The results are shown in Figure 2. The different views of the same object are shown as points of the same colour. From the figure it is clear that the different objects are well separated and form distinct clusters.

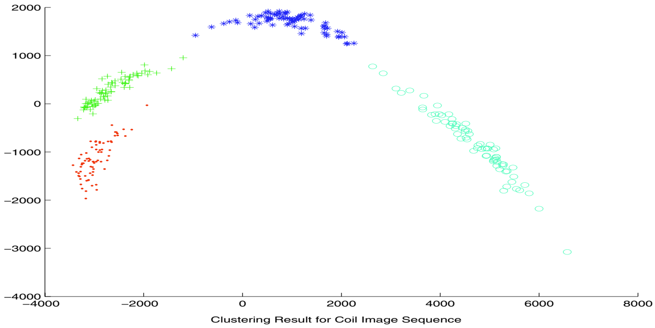


Fig. 2. Clustering Result.

We now turn our attention to the problem of learning the adjacency structure of the graphs. For this experiment we first take 40 sequential images from the "duck" object and 10 images from the "cup" object from the COIL database. By converting the graphs into strings, and applying the graph matching method, we construct the edit distance matrix shown in the left-hand panel of Figure 3. Then the reference string is selected which will generate the minimum edit distance to the set of strings. We then reconstruct the adjacency matrix in a standard order for each graph according to their correspondences to the reference string. After transforming the adjacency matrices into long vectors, we compute the covariance matrix from the set of long vectors. By applying the principal component analysis(PCA), we obtain the two dimensional eigenspace shown in the centre-panel of the figure. Here the "circle" symbols denote the duck image sequence and the "plus" symbols denote the cup sequence. To take this analysis one

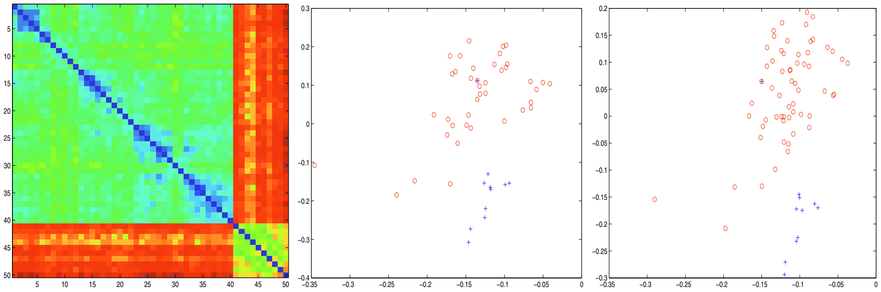


Fig. 3. Edit Distance matrix (left), results of PCA (right).

step further, we add 20 more images from the duck sequence to the set described above. The result of repeating the analysis is shown in the right-hand panel of the figure. From these results, we can see that, under the guidance of the reference string, the two cluster of the objects are well separated.

7 Conclusions

In this paper we have shown how graphs can be converted to strings using semi-definite programming. This is convex optimisation procedure that uses randomised hyperplanes to locate the solution. We have used the resulting strings for the purposes of clustering and analysing the nodes of structural variation for sets of graphs. The graph clusters produced by the method are well separated and the stings delivered by the method can be used to capture the modes of variation in the structure of graphs of a particular class.

References

1. A.Robles-Kelly and E.R.Hancock. Graph Edit Distance from Spectral Seriation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, To appear, 2004.
2. C.Schellewald and C.Schnőrr. Subgraph Matching with Semidefinite Programming . *Proceedings IWCIA (International Workshop on Combinatorial Image Analysis), Palermo, Italy*, 2003.
3. Henry Wolkowicz and Qing Zhao. Semidefinite Programming relaxation for the graph partitioning problem. *Discrete Appl. Math*, pages 461–479, 1999.
4. J.Keuchel,C.Schnőrr,C.Schellewald,and D.Cremers. Binary Partitioning, Perceptual Grouping, and Restoration with Semidefinite Programming. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(11):1364–1379, 2003.
5. Jonathan E. Atkins, Erik G. Boman and Bruce Hendrickson. A Spectral Algorithm for Seriation and the Consecutive Ones Problem. *SIAM Journal on Computing*, 28(1):297–310, 1998.
6. K.Fujisawa,Y.Futakata,M.Kojima,K.Nakata and M.Yamashita. Sdpa-m user's manual. *http://sdpa.is.titech.ac.jp/SDPA-M*.
7. L.Vandenberghe and S.Boyd. Semidefinite Programming. *SIAM Review*, 38(1):49–95, 1996.
8. M.Veldhorst. Approximation of the consecutive one matrix augmentation problem. *J.Comput.*, 14:709–729, 1985.
9. M.X.Goemans and D.P.WIlliamson. Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming. *J.ACM*, 42(6):1115–1145, 1995.
10. W Satish Rao, Andrea and Richa. New approximation techniques for some ordering problems. *ACM-SIAM Symposium on Discrete Algorithms*, pages 211–218, 1998.

Time Reduction of Stochastic Parsing with Stochastic Context-Free Grammars*

Joan Andreu Sánchez and José Miguel Benedí

Depto. Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 Valencia, Spain
{jandreu,jbenedi}@dsic.upv.es

Abstract. This paper proposes an approach to reduce the stochastic parsing time with stochastic context-free grammars. The basic idea consists of storing a set of precomputed problems. These precomputed problems are obtained off line from a training corpus or they are computed on line from a test corpus. In this work, experiments with the UPenn Treebank are reported in order to show the performance of both alternatives.

1 Introduction

Stochastic Context-Free Grammars (SCFGs) are an important specification formalism that are frequently used in Syntactic Pattern Recognition. SCFGs have been widely used to characterize the probabilistic modeling of language in Computational Linguistics [1, 3, 9], Speech Recognition and Understanding [6], and Biological Sequence Analysis [8]. An important advantage of this formalism is the capability to model the long-term dependencies established between the different parts of a sentence, and the possibility of incorporating the stochastic information which allows for an adequate modeling of the variability phenomena that are always present in complex problems. A notable obstacle to using these models is the time complexity of the stochastic parsing algorithms that handle them and the algorithms that are used for the probabilistic estimation of the models from a training corpus.

Most of the well-known parsing algorithms are based on the Earley algorithm for SCFGs in General Format [9] or in the Cocke-Younger-Kasami (CYK) algorithm for SCFGs in Chomsky Normal Form (CNF) [6]. One of these algorithms for SCFGs in CNF is the *inside* algorithm [4], which allows us to compute the probability of a string given a SCFG by using a Dynamic Programming scheme.

The *inside* algorithm has a time complexity $O(n^3)$ for a string of length n. There are theoretical works that attempt to improve this time complexity. In [10], a version of the CYK algorithm was proposed whose time complexity is $O(M(n))$, where $M(n)$ is the time complexity of the product of two matrices of dimension n. The best known algorithm for multiplying two matrices of dimension n is described in [2], whose time complexity is $O(n^{2.38})$. A similar parsing algorithm could be considered for SCFGs by

* This work has been partially supported by the Spanish MCyT under contract (TIC2002/04103-C03-03) and by *Agencia Valenciana de Ciencia y Tecnología* under contract GRUPOS03/031.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 163–171, 2005.

adequately modifying the *inside* algorithm. However, the large implicit constant associated to this matrix product algorithm could make this modified parsing algorithm only interesting for long strings.

Given these drawbacks, other improvement alternatives should be considered. In this work, we propose a simple technique that allows us to reduce computation time especially for short strings. The basic idea consists of storing a set of precomputed problems associated to short strings. The set of problems can be chosen from a training corpus or it can be composed on line from a test set.

In this work, we explore these two proposals and we report the results of experiments on the UPenn Treebank in order to show the performance of both alternatives.

2 Definitions

A *Context-Free Grammar* (CFG) G is a four-tuple (N, Σ, P, S), where N is a finite set of non-terminal symbols, Σ is a finite set of terminal symbols, P is a finite set of rules, and S is the initial symbol. A CFG is in Chomsky Normal Form (CNF) if the rules are of the form $A \rightarrow BC$ or $A \rightarrow a$ ($A, B, C \in N$ and $a \in \Sigma$). A *Stochastic Context-Free Grammar* (SCFG) G_s is defined as a CFG in which each rule has a probability of application associated to it such that $\forall A \in N$: $\sum_{B,C \in N} \Pr(A \rightarrow BC) + \sum_{a \in \Sigma} \Pr(A \rightarrow a) = 1$. We define the *probability* of the derivation d_x of the string x, $\Pr_{G_s}(x, d_x)$ as the product of the probability application function of all the rules used in the derivation d_x. We define the *probability* of the string x as: $\Pr_{G_s}(x) = \sum_{\forall d_x} \Pr_{G_s}(x, d_x)$.

An important problem is the calculation of the probability of a string. For SCFG in CNF, there are different parsing algorithms that are based on the CYK algorithm. We describe one of them below.

The *inside* algorithm [4] allows us to compute the probability of a string by defining $e(A < i, i+l >) = \Pr_{G_s}(A \overset{*}{\Rightarrow} x_i \cdots x_{i+l})$, $0 \leq l < n$, as the probability of the substring $x_i \ldots x_{i+l}$ being generated from A. This probability can be efficiently computed for a string of size n with the following Dynamic Programing scheme for all $A \in N$:

$$e(A < i, i >) = \Pr(A \rightarrow x_i) \quad 1 \leq i \leq n,$$

$$e(A < i, i+l >) = \sum_{\substack{B,C \in N: \\ (A \rightarrow BC) \in P}} \sum_{k=i}^{i+l-1} \Pr(A \rightarrow BC) e(B < i, k >) e(C < k+1, i+l >) \quad (1)$$

$$1 \leq l < n, \; 1 \leq i \leq n-l.$$

In this way, $\Pr_{G_s}(x) = e(A < 1, n >)$.

First, we analyze the time complexity of the *inside* algorithm from expression 1. In the next section, we explain how the computation time can be improved.

Note that the inner loop in the *inside* algorithm comprises two products and one addition. Suppose that we denote with the two products and the addition by a. Then,the total amount of operations is:

$$\sum_{l=1}^{n-1} \sum_{i=1}^{n-l} \sum_{k=i}^{i+l-1} a|P| = \frac{n^3 - 3n^2 + 2n}{3} a|P|. \quad (2)$$

Consequently, the time complexity of the *inside* algorithm is $O(n^3|P|)$.

3 Time Reduction of the Stochastic Parsing Algorithm

Here, we explain how reduce the time required for the stochastic parsing. We state the improvement that can be obtained and finally, we explain the main disadvantage of the proposal.

Note that in expression (1), each substring is a subproblem in the Dynamic Programing scheme. One possible way to reduce computations in expression (1) consists of precomputing all the problems. In this case, expression (1) can be computed by consulting such precomputed problems.

It should be pointed out that with this proposal, the efficient search of a precomputed problem becomes an serious problem. In order to carry out this search efficiently, we have used hash tables. By using this data structure, the search time can be done linearly with the length of the subproblem.

If the time complexity of looking for a subproblem of length l is l times c, where c is the implicit constant associated to the search in the hash table, then (2) becomes $cf(n)$, where $f(n) = (n^3 - 3n^2 + 2n)/3$. Note that for real tasks it is reasonable to think that $c << a|P|$ since $|P|$ can be large.

Note that, for real tasks, it is not feasible to have precomputed all the subproblems due to the amount of memory required. However, it is feasible to have precomputed all the subproblems associated to short strings. For example, if we suppose for simplicity that $|P| = 1$, $a = 3c$, and that we have precomputed all the subproblems up to some size $l \leq n$, we can then save:

$$\frac{f(n) - f(l) + f(l)/3}{f(n)} = 1 - \frac{2f(l)}{3f(n)} = g(l, n).$$

In Figure 1, we have plotted function $g(l, n)$ times 100 for some values of l. This figure shows the savings in percentage depending on the string length. It even shows savings for small values of l. Note that the assumptions have been simplified, and therefore, even more savings can be obtained for a real task.

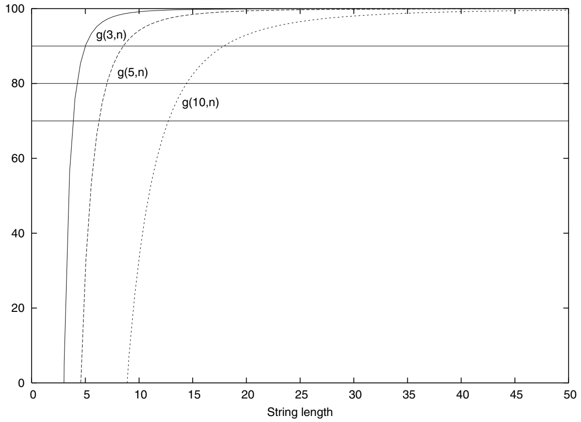


Fig. 1. Percentage of saving depending on the string length.

The disadvantage of this the amount of memory that is needed to store the subproblems. The amount of memory to store a subproblem is linear with the number of non-terminal symbols, and the number of subproblems growths exponentially with the size of the subproblems. Therefore, a trade-off between the amount of memory required and the amount of computation savings must be established.

Given that it is not realistic to store all the subproblems computed, in this work, we propose computing only those subproblems that appear in a corpus. The subproblems can be obtained off line from a training corpus or on line from the test set. In the following section, we study these proposals and we explore practical alternatives for improving the time complexity without increasing the required memory.

4 Experiments

In this section, we describe the experiments that were carried out to test the alternatives proposed in Section 3.

The corpus used in the experiments was the part of the Wall Street Journal that had been processed in the UPenn Treebank project [5]. It contains approximately one million words distributed in 25 directories. This corpus was automatically labeled, analyzed and manually checked as described in [5]. There are two kinds of labeling: a POStag labeling and a syntactic labeling that is represented by brackets. The POStag vocabulary is composed of 45 labels; and the syntactic vocabulary is composed of 14 labels. The corpus was divided into sentences according to the bracketing and, following other works, sentences with more than 50 words were ignored (this represented less than 2% of the corpus). For the experiments, the corpus was divided into two sets: training (directories 00-20; 41,315 sentences; 959,390 words), and test (directories 23-24; 3,702 sentences; 86,053 words).

Given that we needed a SCFG for the stochastic parsing, we took advantage of an SCFG that had been estimated in a previous work [1]. This SCFG was learned with sentences labeled with POStags as described in [1]. The estimation algorithm was the bracketed version of the *inside-outside* algorithm [1, 7]. The final estimated SCFG had 35 non-terminal symbols, 45 terminal symbols, and 1,741 rules.

Here, we describe the experiments that were carried out to test the proposed technique. Hash tables were used to store the subproblems in all the experiments. All the software was implemented in C language and the gcc compiler (version 3.3.1) was used. All experiments were carried out on a personal computer with an Intel Pentium 4 processor of 2.40 GHz, with 1.5 GB of RAM and with a Linux 2.4.21 operating system.

In all the figures, we have plotted the percentage of computation of the proposed technique with respect to the unmodified *inside* algorithm when parsing the test set. In order to evaluate the influence of the length of the parsed strings, the test set was gradually enlarged by incorporating strings of increasing size.

First, we present experiments in which the subproblems were obtained off line from a training corpus. Then, we present experiments in which the subproblems were obtained on line from the test set.

4.1 Experiments with Subproblems Obtained from the Training Set

In this section, we present three experiments. In the first experiment, all the subproblems that appeared in the training corpus were stored. In the second and third experiments, we studied techniques to reduce the required memory.

Experiment with All the Subproblem of the Training Set. In this experiment, we tested the proposed technique without memory restrictions. We obtained all the subproblems from the training set up to a given size. Table 1 shows the number of subproblems for different sizes (probabilities are represented with integers of four bytes) and the amount of accumulated memory that was necessary to store all the subproblem up to a given size. Note that the amount of memory increased notably as the size of the stored subproblems increased.

Table 1. Number of subproblems in the training set and the required memory.

Subproblem size	2	3	4	5	6
No. of subproblems	1,353	15,276	75,702	209,239	377,254
Accumulated memory (in MB)	0.18	2.22	12.33	40.27	90.64

Figure 2 shows the percentage of computation time with respect to the unmodified *inside* algorithm with the test set. In the x-axis, we represent the maximum length of sentences content in the test set. Therefore, each point in the curve stands for the percentage of computation time reduction when strings of the test set up to a given length were parsed.

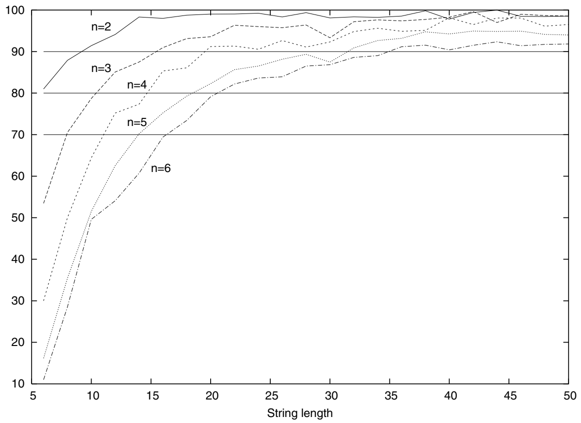


Fig. 2. Percentage of reduction in computation time with the test set.

In this figure, it is important to point out two issues. First, the computation saving improved as expected as the size of stored subproblems increased. Second, the computation saving that was obtained for short strings was very good. Thus, when subproblems up to length six were stored, a computation savings of almost 35% was obtained for strings of length 15.

Note that not all problems in the training set have to be equally frequent and some of them might even be unnecessary.

Selection of the Subproblem According to the Frequency of Occurrence. In this experiment, we evaluated our proposal when storing only those subproblems that occurred at least twice in the training set for the subproblems of largest length (5 and 6). Table 2 shows seen the number of subproblems for different lengths and the amount of memory required to store them. We can see that the amount of required memory decreased notably.

Table 2. Number of subproblems in the training set when the less frequent problems (of length 5 and 6) were removed and the required memory.

Subproblem size	2	3	4	5	6
No. of subproblems	1,353	15,276	75,702	81,823	94,084
Accumulated memory (in MB)	0.18	2.22	12.33	23.26	35.82

Figure 3 shows the percentage of reduction in computation time when this idea was used (curve `e2`). Curve `e1` corresponds to curve `n=6` in Fig. 2. The computation savings was unaffected by this proposal. It should be pointed out, that even though the amount of required memory decreased notably, problems that are not very probable according to the grammar might be stored.

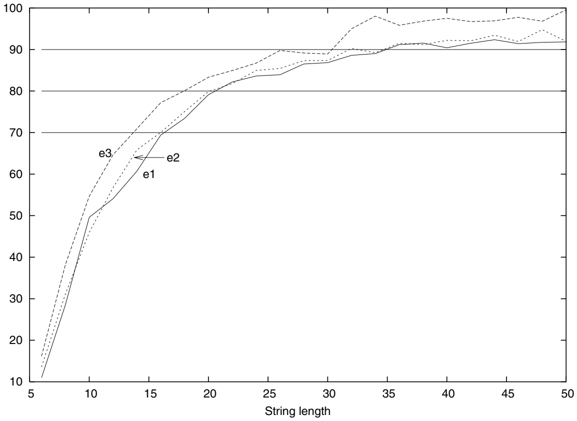


Fig. 3. Percentage of reduction in computation time with the test set when some subproblems were removed.

Selection of the Subproblems According to Their Parsing Probability. Another criterion for choosing the set of subproblems could be based on the parsing probability.

An experiment was carried out in which the most "probable" subproblems of sizes 5 and 6 of the training set were stored. In order to choose the most "probable" subproblems, the following function was used: $\sum_{A \in N} \mathrm{Pr}_{G_s}(A \overset{*}{\Rightarrow} x_1 \cdots x_l)$, where N is the set of non-terminal symbol of the SCFG. This function can be interpreted as the average probability of a subproblem of size l being generated from a non-terminal symbol. From this function, the following threshold was defined: $(\sum_{p \in P_l} \sum_{A \in N} \mathrm{Pr}_{G_s}(A \overset{*}{\Rightarrow} x_1 \cdots x_l))/|P_l|$, where P_l is the set of subproblems of size l. In this experiment, those subproblems of sizes 5 and 6 that did not overcome this threshold were removed. A notable reduction in memory consumption was achieved this way (see Table 3).

Table 3. Number of subproblems in the training set when the less "probable" problems of length 5 and 6 were removed and the required memory.

Subproblem size	2	3	4	5	6
No. of subproblems	935	6,244	17,708	25,861	52,098
Accumulated memory (in MB)	0.13	0.96	3.32	6.77	13.73

Figure 3 (curve `e3`) shows that the computation savings was slightly worse.

4.2 Experiments with Subproblems Obtained On-Line from the Test Set

One important problem that can appear when subproblems are stored is that the distribution of strings in the training and test can be very different. Therefore, the stored problems may not be useful. One possible solution to overcome this issue is to store on line only those problems that appear in the test set. Note that in this case, additional time consumption is required. Given that we store subproblems on line, the memory consumption increases as the test set is being parsed.

We tested this proposal and Table 4 shows the memory required for the test set when all the sentences were parsed. Note that this memory consumption was obtained when sentences up to size 50 were parsed, that is at the end of the experiment.

Table 4. Number of subproblems in the test set and the required memory.

Subproblem size	2	3	4	5	6
No. of subproblems	984	6,964	21,691	40,075	53,019
Accumulated memory (in MB)	0.13	1.06	3.96	9.31	16.39

Figure 4 shows the computation savings in this experiment. Curve `e1` corresponds to curve `n=6` in Fig. 2. The savings was worse for long strings than in the experiments described in Section 4.1. This could be due to the fact that the time used to store the subproblems was included in the computed time.

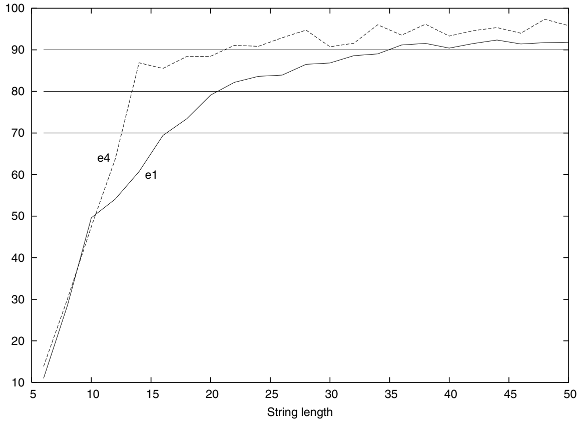


Fig. 4. Percentage of reduction in computation time with the test set when some subproblems were removed.

4.3 Experiments with Subproblems Obtained Off-Line from the Training Set and On-Line from the Test Set

Note that both the ideas from Section 4.1 and 4.2 can be combined; that is, we could use subproblems obtained off line from the training set and subproblems obtained on line from the test set. In the final experiment, we used both techniques to obtain the precomputed subproblems.

In order to avoid excessive memory consumption, we first combined the selection of the subproblem according to the frequency of occurrence in the training set with the subproblems obtained on line from the test set (see Fig. 5 curve `e2,e4`). We then combined the selection of the subproblems of the training set according to their parsing probability with the subproblems obtained on line from the test set (see Fig. 5 curve `e3,e4`).

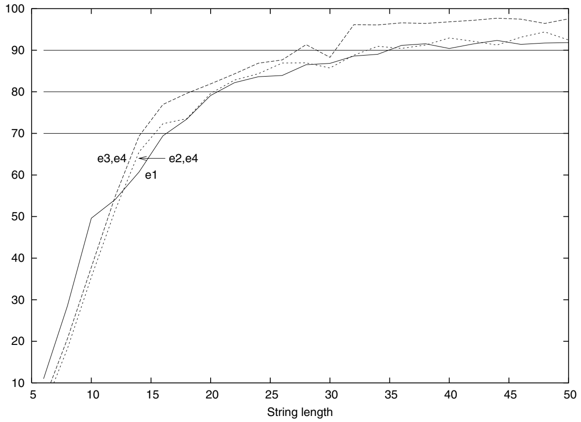


Fig. 5. Percentage of reduction in computation time with the test set.

This combined technique provided good results for short strings (below 13). Above 13, the best results were obtained when the selection of subproblems from the training set off line was carried out according to frequency of occurrence.

5 Conclusions

A novel technique has been introduced to save computation time using the *inside* parsing algorithm. The basic idea is to store precomputed problems and use them in the parsing algorithm. These precomputed problems can be obtained off line from a training corpus or on line from the test set. This technique was successfully applied in a real experiment and an important reduction in computation time was achieved. This reduction was more accentuated in short strings. This fact is especially important for tasks like Automatic Speech Recognition, where sentences are usually short.

For future work we plan to apply this technique to tasks where the vocabulary size is small such as RNA Sequence Modeling. Another possibility is to apply this technique to the *inside-outside* estimation algorithm for SCFG.

References

1. J.M. Benedí and J.A. Sánchez. Estimation of stochastic context-free grammars and their use as language models. *Computer Speech and Language*, 2005. To appear.
2. D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *J. Symb. Comput.*, 9(3):251–280, 1990.
3. F. Jelinek and J.D. Lafferty. Computation of the probability of initial substring generation by stochastic context-free grammars. *Computational Linguistics*, 17(3):315–323, 1991.
4. K. Lari and S.J. Young. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 4:35–56, 1990.
5. M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.
6. H. Ney. Stochastic grammars and pattern recognition. In P. Laface and R. De Mori, editors, *Speech Recognition and Understanding. Recent Advances*, pages 319–344. Springer-Verlag, 1992.
7. F. Pereira and Y. Schabes. Inside-outside reestimation from partially bracketed corpora. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 128–135. University of Delaware, 1992.
8. I. Salvador and J.M. Benedí. Rna modeling by combining stochastic context-free grammars and n-gram models. *International Journal of Pattern Recognition and Artificial Intelligence*, 16(3):309–315, 2002.
9. A. Stolcke. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21(2):165–200, 1995.
10. L.G. Valiant. General context-free recognition in less than cubic time. *Journal of computer and system sciences*, 10:308–315, 1975.

Part III

Image Analysis

Segment Extraction Using Burns Principles in a Pseudo-color Fuzzy Hough Transform

Marta Penas[1], María J. Carreira[2],
Manuel G. Penedo[1], and Cástor Mariño[1]

[1] Dpto. Computación, Fac. Informática, Universidade da Coruña,
15071 A Coruña, Spain
{infmpc00,cipenedo,castormp}@dc.fi.udc.es

[2] Dpto. Electrónica e Computación, Universidade de Santiago de Compostela,
15782 Santiago de Compostela, Spain
mjose@dec.usc.es

Abstract. This paper describes a computational framework developed for the extraction of low-level directional primitives present in an image, and subsequent organization through a line segment detector. The system is divided in three stages: extraction of the directional features in the image through an efficient implementation of Gabor wavelet decomposition; reduction of these high dimensionality results by means of a growing cell structure; and extraction of the segments from the image. This last step was first implemented through a pseudo-color Fuzzy Hough Transform and then improved through some principles of the Burns segment detector.

Keywords: Gabor wavelets, growing cell structures, chromaticity diagram, Hough transform, Burns segment detector.

1 Introduction

The boundaries of objects in an image often lead to oriented and localized changes in intensity called edges. Edge and segment detection are the first steps in many image analysis applications and they are of great importance as they constitute the basis for the higher levels in the system. It has always been a fundamental problem in computer vision that the higher level processing stages suffer due to either too little or too much data from the lower levels of the processing. Thus, the quality of data available for further analysis is very critical.

This paper describes a framework for the extraction of the directional properties present in an image through Gabor wavelet decomposition and the detection of the segments that approximate these properties through a segment detector based on the fuzzy Hough transform and the Burns segment detector.

The Gabor Wavelet decomposition framework presented here is a computationally expensive process, but provides precise information about the orientation of image pixels and is independent of image type. Moreover, we have implemented [1] an approximation to Gabor Wavelets that reduces the computational time and memory requirements through the use of a Gabor Wavelet

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 175–182, 2005.

decomposition in the spatial domain, which is faster than conventional frequency domain implementations.

A previous paper [2] describes the use of a pseudo-color fuzzy Hough transform for the detection of the segments present in the image. The pseudo-color Hough transform works properly in simple images that only contain the objects to detect, but has some limitations in complex images due to its global nature. This is the reason why, in this paper, we propose a refinement of the pseudo-color Hough transform through the introduction of some principles of the Burns segment detector.

This paper is organized as follows. Sec. 2 describes the extraction of directional primitives present in the image, Sec. 3 describes the segment extraction process through a pseudo-color fuzzy Hough transform and Sec. 4 describes the introduction of the Burns segment detector principles. Finally, Sec. 5 contains the conclusions from our work.

2 Extraction of Directional Primitives

This section contains a brief introduction of the first stages of the process, the extraction of the directional primitives present in the image through Gabor wavelet decomposition and the organization of these results through a growing cell structure. These stages construct the RGB images that the segment detector developed in this work will receive as input.

Gabor wavelets [3] are complex exponential signals modulated by Gaussians with two important properties that make them good edge detectors: the optimization of edge localization [4] and the absence of image-dependent parameter tuning. Their most important drawback is their greedy demand for both memory and computation time. In a previous paper [2], we developed a more efficient, multi-resolution spatial domain implementation of Gabor Wavelet decomposition, which we employ here, based on the convolution of 11 1D-component masks obtained through the decomposition of the 2D masks that define the wavelets. The implementation here uses the good edge localization property of Gabor wavelets, with the exact position of an edge determined as a conjunction between a maximum in the modulus and a zero crossing in the even or the odd part of Gabor results.

In our Gabor decomposition, the input image is filtered with a bank of 8 filters centered at frequency $\frac{1}{4}$ and 8 orientations ($\frac{k\pi}{8}, k = 0..7$) leading to 8 resulting images. A reduction of this output space dimensionality is necessary in the interest of efficiency. Auto-organized structures are a suitable instrument to achieve this dimensionality reduction as they allow simultaneously the reduction of the input space and the projection of the topological order in the input space to the output structure.

In [5], self-organized maps, growing cell structures and growing neural gas structures were investigated and compared for their power of dimensionality reduction of Gabor decomposition results. Growing cell structures (GCS) [6] provided significantly better results. They are artificial neural networks based on self-organized maps that eliminate the restrictions of the *a priori* network size definition, incorporating a mechanism to add new processing elements when needed, while maintaining the network topology.

To represent the different directionalities provided by the auto-organized structure, each processing element was assigned a color from a colormap to indicate its orientation. The colormap was obtained from 8 equidistant points on the perimeter of the maximum circle inside the RGB triangle in the chromaticity diagram [7], centered at white (see Fig. 1-left). Fig. 1 right shows the GCS output from a ring demonstrating the colors of the entire direction space, i.e. $0 - 2\pi$.

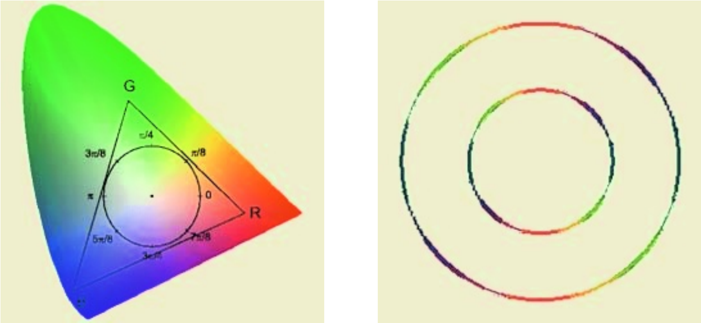


Fig. 1. Left: Colormap circle inside the RGB triangle. Right: All orientations after the GCS analysis.

Figure 2 shows the results from three input images. The first row shows the original images and second row the results from GCS analysis. Fig. 2 left shows a medical image that contains protein crystals exhibiting polygonal shapes, Fig. 2 center shows an IR aerial image of a bridge and Fig. 2 right is the picture of a boat.

3 Segment Extraction Through the Fuzzy Hough Transform

The Hough transform is widely used in artificial vision and pattern recognition for the detection of geometrical shapes that can be defined through parametric equations. Traditional Hough transform implementations are based on the results obtained by classical edge detectors, like Canny or Sobel. We have designed and implemented a pseudo-color fuzzy Hough transform [2] based on the pseudo-color images obtained from the processes described in Sec. 2, that is, color images where each color represents a specific orientation.

The Hough transform is based on the normal equation of a line, which states that, if the normal to the line makes an angle θ with the x axis and the length of this normal is ρ, then the equation of the line is given by:

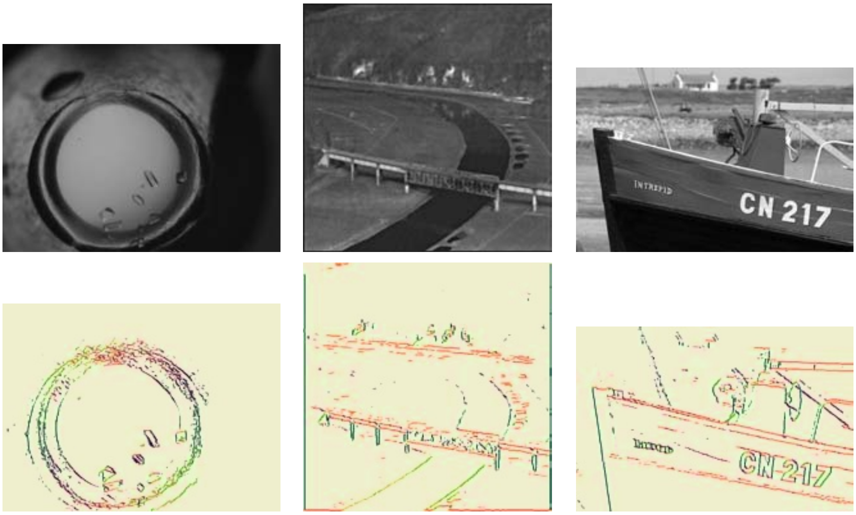


Fig. 2. First row: images 'proteins', 'bridge' and 'boat'. Second row: results of GCS analysis.

$$\rho = x \cdot \cos\theta + y \cdot \sin\theta \tag{1}$$

Classically, the continuous $\rho - \theta$ space is quantized into suitable sized $(\triangle\theta \times \triangle\rho)$ rectangles and each rectangle is associated with an element of an accumulator array A with size $N_\rho \times N_\theta$. However, this parameterization needs to be enhanced to deal with the lines in our color-labelled images. That is, all points with similar colors (close orientations) and neighboring positions will vote for the same line. Our accumulator will have 2 dimensions $p = 0..N_\rho - 1$ and $q = 0..N_\theta - 1$, and we must search for the maxima in this space.

The process begins with the quantization of the Hough space in $N_\rho \times N_\theta$ cells, where $N_\rho = \sqrt{I_x^2 + I_y^2} / \triangle\, \rho$ and $N_\theta = \pi / \triangle\, \theta$, depending on the size of the image $(I_x \times I_y)$ and the quantization in ρ $(\triangle\rho)$ and θ $(\triangle\theta)$. Then, we compute the contribution of each pixel P in the labelled image, from its angle q_P, determined by the color it has assigned. For this angle, we compute the corresponding θ_P from the quantization in θ $(\theta_P = q_P \triangle\, \theta)$, and then ρ_P from eq. 1. Then, the quantized value p_P is obtained from ρ by $(p_P = \rho_P / \triangle\, \rho)$. A pixel P contributes to all the neighbors $A(p, q)$ of $A(p_P, q_P)$ incrementing their value with $\triangle A = C_p C_q$ where:

$$C_p = e^{-\beta_p d^2(p, p_P)} \text{ and } C_q = e^{-\beta_q d^2(q, q_P)} \tag{2}$$

p being the p-neighbor of p_P if $C_p < \varepsilon_p$ and q the q-neighbor of q_P if $C_q < \varepsilon_q$. β_p and β_q are the parameters of the Gaussians that define the contributions to the accumulator $\mathbf{A}$, and d is the distance between the two points in p-space and q-space, respectively. These parameters control how smooth the decay in the Gaussian is.

Parameters β_p and ε_p are fixed from $\triangle\rho$. As we have chosen $\triangle\rho = 1$, we have fixed $\beta_p = 0.1$ and $\varepsilon_p = 0.4$. With these β_p and ε_p, when the distance between p and p_P is greater than about 4, the contribution is too small to be considered. Parameters β_q and ε_q were selected from $\triangle\theta$ similarly. First, we have fixed $\triangle\theta = \pi/24$. As C_q must have a high contribution to the nearest angles not overlapping the previous and next main orientations, we have fixed $\beta_q = 0.5$ and $\varepsilon_q = 0.1$ and so the contributions of angle indices q further than 3 from q_P are very small.

Once the voting process has finished, the following step is the maxima detection. Each maximum detected in the accumulator array corresponds to a line in the image that can contain one or more segments. For each maximum detected over a predefined threshold, an inverse Hough transform removes all the contributions to the accumulator array of the pixels belonging to the line detected. A segment detection takes place simultaneously to the inverse Hough transform. When a pixel belonging to a line is removed from the accumulator array, its orthogonal projection to the line is determined. Once all the pixels involved have been analyzed, the line is sequentially searched to determine which pixels belong to each segment.

The final result of this process is an array of segments. Each segment is defined by the polar coordinates of the line it belongs to and its endpoints. From these, all the defining characteristics of the segment, like length or slope, can be computed.

Fig. 3 shows the results from segment detection through the pseudo-color fuzzy Hough transform applied to the images in the first row of Figure 2. As the Hough transform is a global instead of a local operation, the information contained in the whole image can influence the deviation of the segments and the detection of spurious segments in the result images as Fig. 3 shows. This is the reason why we have implemented a new segment detector based on the combination of our pseudo-color Hough transform and some principles of the Burns segment detector.

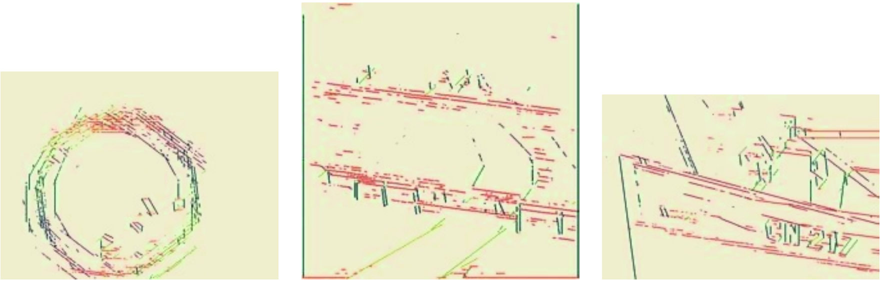


Fig. 3. Results of segment detection through our pseudo-color Hough transform from images in first row of Fig. 2.

4 Burns Segment Detector and Pseudo-color Fuzzy Hough Transform

As previously mentioned, the Hough transform is a global process where each pixel votes for many possible lines. Opposite to this, the Burns segment detector [8] organizes globally the supporting line context prior to any decision. The approach consists of grouping image pixels across the width and the length of the edge through their orientation to form a line support region. Then, only the pixels in the line-support region contribute to the final representation of the line.

The basic steps of the segment detector are: the grouping of pixels into line-support regions based on their orientation and the application of the pseudo-color fuzzy Hough transform described in the previous section to each line-support region separately.

In order to group pixels into line-support regions, the π radians range of orientations is quantized into 4 angular partitions of $\frac{\pi}{4}$ radians starting at 0. Then, each edge pixel is labeled according to the partition into which its orientation falls. If our estimation of the orientation is correct, the pixels belonging to a line will belong to the same partition or sometimes to adjacent partitions if their orientation is close to one of the partition boundaries. The simple connected-components algorithm is then used to form distinct regions for groups of adjacent pixels lying in the same angular partitions.

If we use only one partition of the orientation range, two problems can arise. First, two contiguous lines can improperly be merged if they have similar orientations that lie in the same angular interval. Second, the lines lying across a partition boundary could produce fragmented support regions. The over-merging problem tends to be reduced as the partition size gets smaller, but then the fragmentation problem gets worse.

In the proposed line segment detector, the over-merging problem is solved by the pseudo-color fuzzy Hough transform. As it uses a finer quantization of the orientation range (partitions of size $\frac{\pi}{24}$) it can separate contiguous lines improperly merged into a unique line support region.

In order to solve the fragmentation problem, a new set of partitions is introduced. This partition overlaps the previous one and divides the π radians range of orientations into 4 intervals of $\frac{\pi}{4}$ radians starting at $\frac{\pi}{8}$. Each edge pixel is again labeled according to its orientation. When a set of partitions fragments a line that lies across a boundary, the other set will place it in the same partition.

Both sets of partitions must be merged in such a way that each pixel is associated to only one line-support region. The region considered best for the pixel is the one that provides an interpretation of the line that is the longest.

Each line-support region represents a candidate area for a straight line or some contiguous straight lines with similar orientations. So, the pseudo-color fuzzy Hough transform described in the previous section is applied in order to detect the line segments that it contains. Results obtained from the segment detector just described applied to the images in the first row of Fig. 2 are shown in Fig. 4. As in the previous section, the final output of the process is an array of segments, each segment is defined by its polar coordinates and its endpoints.

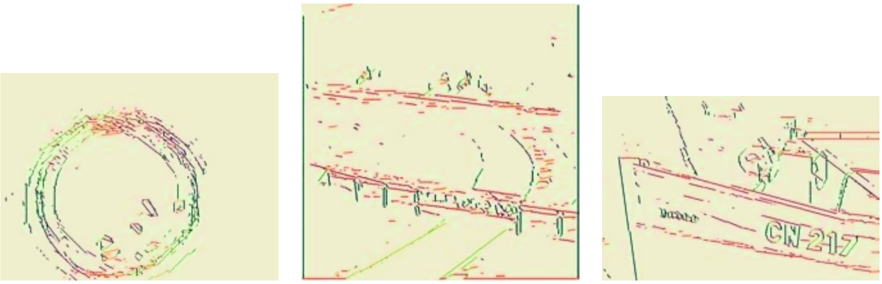


Fig. 4. Results of segment detection through the proposed segment detector from images in first row of Fig. 2.

As Figure 4 shows, the segment detector just described improves the results obtained by the segment detector described in Sec. 3 because the segments detected approximate more accurately the edges of the underlying scene. As a practical example, the image shown in first column of Fig. 2 contains protein crystals exhibiting polygonal shapes. In an application developed for the detection of such crystals in images similar to this, the results obtained from the segments detected by our last segment detector would be much more accurate than the results from the segments detected by our previous segment detector, as the images in Fig. 5 show. These images correspond to a zoom over some area of the original image shown in Figure 2 that contains a polygonal shape.

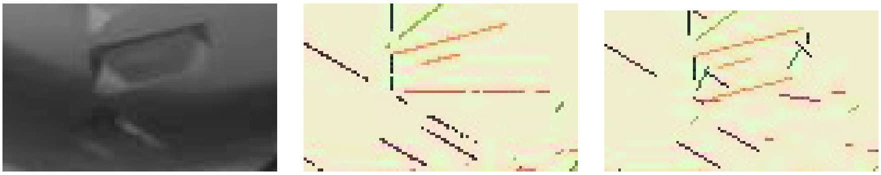


Fig. 5. Left: zoom over original image in Fig. 2. Center: segment detection results by the pseudocolor fuzzy Hough transform. Right: results of segment detection through the proposed segment detector.

5 Conclusions and Discussion

This paper describes a computational framework for the detection of line segments in 2D images. Its first stage consists of the extraction of the directional primitives through Gabor wavelet decomposition. The second stage consists of the organization of these low-level directives through growing cell structures. And the third stage consists of the segment detection through a combination of principles from the Burns segment detector and the fuzzy Hough transform.

First, a novel implementation of the fuzzy Hough transform was developed. This fuzzy Hough transform works with the pseudo-color images provided by the previous stages of the process and its output is the list of segments present in the input scene. The main limitation of the fuzzy Hough transform is its global nature and causes the deviation of segments and generation of spurious segments. As this paper shows, this limitation can be overcomed through the application of some principles of the Burns segment detector, that discretizes the image into line-support regions.

Segment detection results are the basic primitives for a wide range of image processing techniques, like object detection. Improving the results from segment detection has a great influence over the quality of the final results. A practical example is shown in Figure 3 and Figure 4. Detecting the objects present in the original images shown in Fig. 2 is easier from the segments detected by the Burns segment detector that from the segments detected by the fuzzy Hough transform, as these segments approximate more accurately the edges in the original images.

The edge detectors just described have been tested over a wide range of images. Some of them can be accesed at `http://www.lfcia.org/~marta/IbPRIA2005`.

Acknowledgements

We gratefully acknowledge the support of this research by Prof. Linda Shapiro and the NIH National Institute of General Medical Sciences, Grant Number 1 R01 GM068878. This paper has also been partly supported by the University of A Coruña and by the Xunta de Galicia through grant PGiDT03TIC10503PR.

References

1. M. Penas, M. J. Carreira and M. G. Penedo. Gabor wavelets and auto-organised structures for directional primitive extraction. Lect. Notes Comp. Science **2652** (2003) 722–732.
2. M. Penas, M. J. Carreira and M. G. Penedo. Perceptual organization of directional primitives using a pseudocolor Hough transform. Lect. Notes Comp. Science **2749** (2003) 893–898.
3. B. Gabor. Theory of Communication. Journal of the Institute of Electronic Engineers **36(93)** (1946) 429–457.
4. J. Van Deemter and J. Du Buf. Simultaneous Detection of Lines and Edges Using Compound Gabor Filters. Journal of Pattern Recognition and Artificial Intelligence **14(4)** (2000) 757–777.
5. M. Penas, M. J. Carreira and M. G. Penedo. Auto-organised structures for extraction of perceptual primitives. Lect. Notes Comp. Science **2085** (2001) 628–636.
6. B. Fritzke. Growing Cell Structures - A Self-organizing Network for Unsupervised and Supervised Learning. Neural Networks **7(9)** (1994) 1441–1460.
7. G. Wyszecki and W. S. Stiles. Color science, concept and methods, quantitative data and formulae. John Wiley & sons (1982).
8. J. B. Burns, A. R. Hanson and E. M. Riseman. Extracting Straight Lines. IEEE Trans. on Pattern Analysis and Machine Intelligence **PAMI-8(4)** (1986) 425–455.

Texture Interpolation Using Ordinary Kriging

Sunil Chandra, Maria Petrou, and Roberta Piroddi

Centre for Vision, Speech and Signal Processing
School of Electronics and Physical Sciences
University of Surrey, Guildford GU2 7XH, UK
{s.chandra,m.petrou,r.piroddi}@surrey.ac.uk

Abstract. We present a survey of the application of ordinary Kriging to texture interpolation using a variety of models that have been proposed to model the variogram of the image. The novelty of our approach is in the fully automated process of fitting the models to the data over a finite range of values.

1 Introduction

There are various techniques for interpolating irregularly sampled data [6]. However most of these techniques assume that the missing data possess some sort of smoothness and when they are applied to highly textured images they are not expected to perform well. In this paper we deal with the problem of texture interpolation. This is a particularly difficult problem as texture is a spatial property and irregular sampling may destroy the perceived pattern to a very high degree. Kriging [1],[4] is a method often employed by geoscientists for the interpolation of irregularly sampled data, but it is less well known to the image processing community. In this paper we undertake a thorough study of this methodology and investigate its local as well as global application in conjunction with five models that have been proposed to model the variogram of the image. In section 2 we present the definitions of the various terms relevant to this work, and the methodology we use. In section 4 we present our experimental results and in section 5 we conclude.

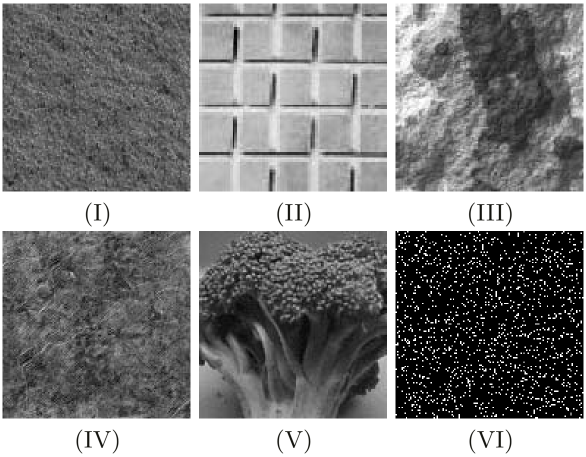
2 Methodology

In general, in order to interpolate using the Kriging method, we must model the covariance matrix of the random variable. This is done by modelling the variogram of the data. There are several variogram models available. In this paper we are using five different variogram models. After the parameters of the variogram model have been identified, we proceed to perform the interpolation.

2.1 Computation of the Variogram

The variogram is defined as the expected squared difference between two data points separated by a distance h. Half of the variogram is known as the semivariogram [3]. It is defined as:

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 183–190, 2005.



(I) (II) (III)

(IV) (V) (VI)

Fig. 1. Test images (I) – (V) and an example sampling mask that retains 9% of the pixels.

$$2\gamma(h) \equiv Var\{V(P_i) - V(P_i + h)\} \equiv E\{[V(P_i) - V(P_i + h)]^2\} \tag{1}$$

where E denotes the expected value operator and $V(P_i)$ denotes the grey level value at point P_i. If the total number of distinct pairs of data points V_i and V_j, whose positions are at a distance $d_{ij} = h$ from each other, is denoted by $N(h)$, then $\gamma(h)$ is also defined as [1]:

$$\gamma(h) = \frac{1}{2|N(h)|} \sum_{(i,j)|d_{ij}=h} (V_i - V_j)^2 \tag{2}$$

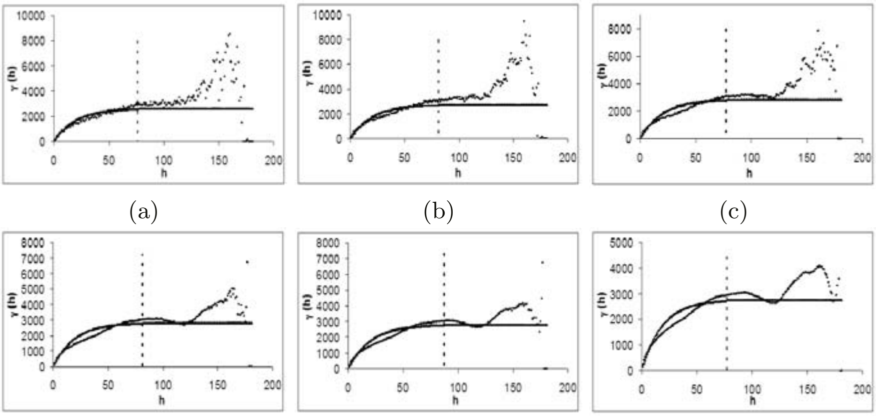
The relationship between the variogram and its corresponding covariance is given by [4],

$$\tilde{C}(h) = \begin{cases} C_0 + C_1 & if \quad |h| = 0 \\ C_0 + C_1 - \gamma(h) & if \quad |h| > 0 \end{cases} \tag{3}$$

where C_0 is the nugget effect and $C_0 + C_1$ is the sill. Although theoretically $\gamma(0) = 0$, in practice $\gamma(0) = C_0$ and this is the nugget effect. Sill the constant value $\gamma(h)$ reaches when it levels off.

2.2 Kriging

Kriging [1],[4] is a popular interpolation method, where the unknown value of the signal $f_0 = f(x, y)$ at a given coordinate position (x_0, y_0) is expressed as a linear combination of the S known, irregularly sampled values of the signal, so that $f(x, y) = \sum_{s=1}^{S} w_s f(x_s, y_s)$, for $s = 1, \ldots, S$. The characteristic of this method is that the weights w_s are chosen in such a way as to minimise the variance of the error made in the estimation of the signal. We define: $\mathbf{W^T} \equiv (\mathbf{w_1}, \quad \ldots, \quad \mathbf{w_S}, \quad \mu)$, where μ is the average of the known samples and $\mathbf{D^T} \equiv$



(a) (b) (c)

(d) (e) (f)

Fig. 2. Local fitting of the exponential variogram model for image III in which only the part of the curve on the left of the dotted line is used for local Kriging. Panels (a) – (e) correspond to 4%, 6%, 9%, 26%, 45% undersampling versions of image III, respectively, as well as to the full image (f).

$\left(\tilde{\mathbf{C}}_{\mathbf{10}}, \quad \tilde{\mathbf{C}}_{\mathbf{20}}, \quad \ldots, \quad \tilde{\mathbf{C}}_{\mathbf{S0}}, \quad \mathbf{1}\right)$, where the elements $\tilde{C_{s0}}$ are the covariances between the known samples and the signal value at coordinate (x_0, y_0), and

$$\mathbf{C} \equiv \begin{pmatrix} \tilde{C}_{11} & \ldots & \tilde{C}_{1S} & 1 \\ \vdots & \ddots & \vdots & 1 \\ \tilde{C}_{S1} & \ldots & \tilde{C}_{SS} & 1 \\ 1 & \ldots & 1 & 0 \end{pmatrix} \tag{4}$$

is the covariance matrix for the known samples. In general, one must model the covariance matrix of a random variable $f(x, y)$. This is done by choosing a covariance function $\tilde{C}$ and calculating all the required covariances from it. The covariance function is obtained by using an appropriate variogram model. The weights are computed according to: $\mathbf{W} = \mathbf{C^{-1}D}$.

3 Fitting the Variogram with the Models

In this section we fit the variogram with five different models by using least square error fitting [5], [2]. The models are fitted globally, to the full variogram, or locally, choosing the range over which they will be fitted, totally automatically. The **fractal model** is defined as [5],

$$\tilde{\gamma}(h) = \gamma_0 h^{2H} \tag{5}$$

where γ_0 is the intercept of the line fitted to the data when we plot $\log(\tilde{\gamma}(h))$ versus $\log h$ as $h \to 0$. In general, the log variogram is not linear for the whole

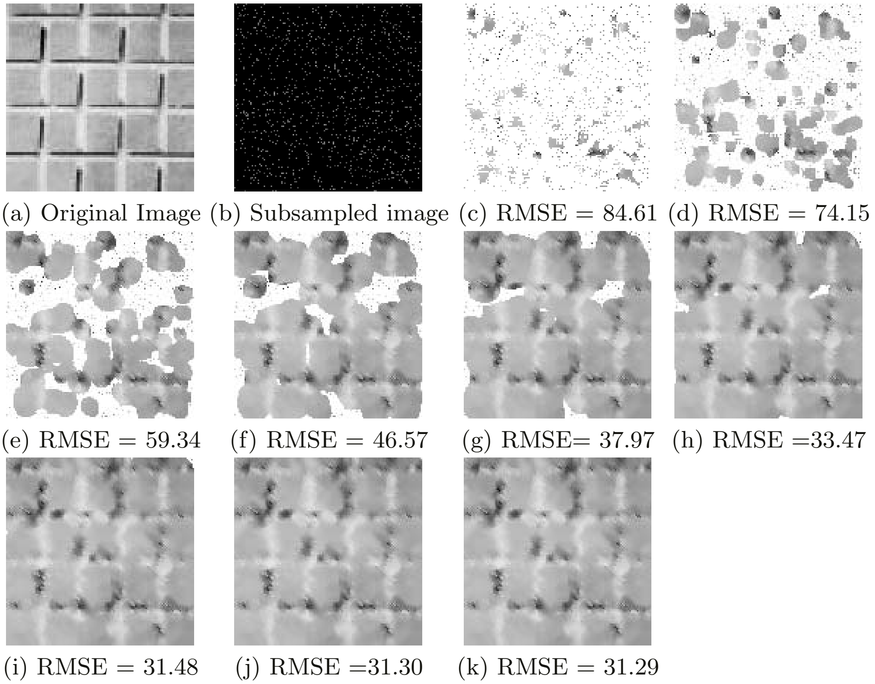


Fig. 3. Process of reconstruction for image II from 6% of its pixels, using the fractal model.

range of h values, so it is obvious that the fractal model cannot possibly be used for the whole range of h values. However, for small values of h the linear model may be applicable. The trouble is that we do not know a priori the range of values of h over which the model may be applied. To solve this problem, we use the correlation coefficient to guide us in choosing the region over which the model fits best, in a recursive way.

The **exponential variogram** model is defined as [1]

$$\tilde{\gamma}(h) = \begin{cases} 0 & \text{if } |h| = 0 \\ C_0 + C_1 \left(1 - \exp\left(\frac{-|h|}{a}\right)\right) & \text{if } |h| > 0 \end{cases} \tag{6}$$

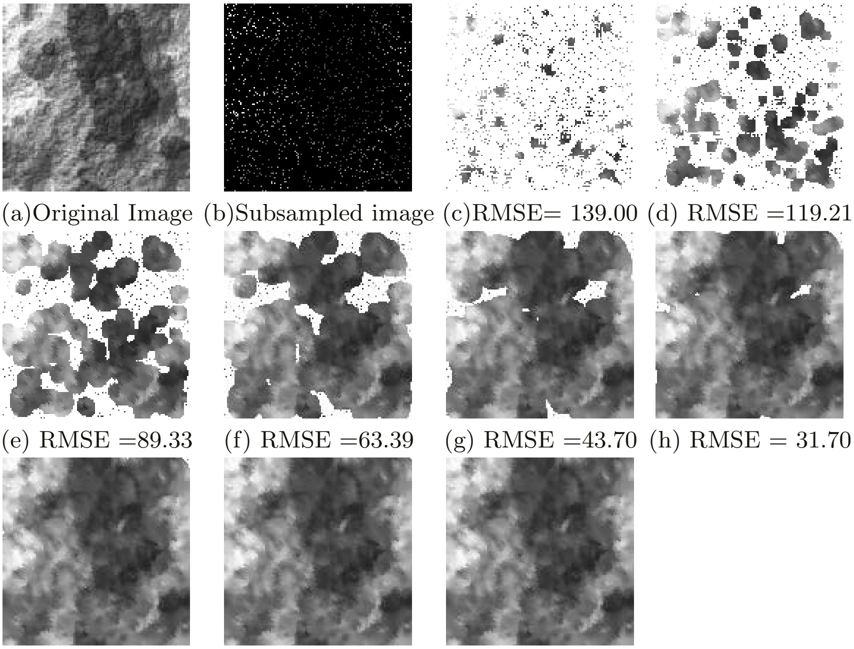
The **spherical model** is defined as [1],

$$\tilde{\gamma}(h) = \begin{cases} C_0 + C_1 & \text{if } |h| \geq a \\ C_0 + C_1 \left(1.5\frac{h}{a} - 0.5(\frac{h}{a})^3\right) & \text{if } |h| < a \end{cases} \tag{7}$$

In both last cases, we first find the range h_{max} of values of h for which $\gamma(h)$ flatters out and then we fit the model for h in the range $[1, h_{max}]$.

The **Gaussian model** is given by [1],

$$\tilde{\gamma}(h) = C_0 + C_1 - C_1 \exp\left(\frac{-|h|^2}{a^2}\right) \tag{8}$$



(a)Original Image (b)Subsampled image (c)RMSE= 139.00 (d) RMSE =119.21

(e) RMSE =89.33 (f) RMSE =63.39 (g) RMSE =43.70 (h) RMSE = 31.70

(i) RMSE = 26.52 (j) RMSE = 21.21 (k) RMSE =21.20

Fig. 4. Process of reconstruction for image III from 6% of its pixels, using the linear model.

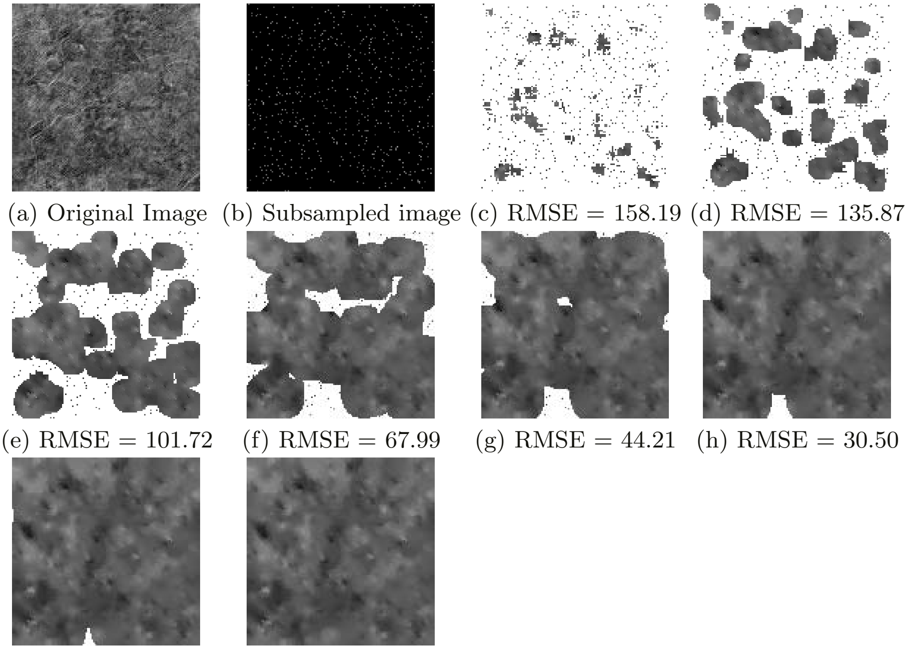
This model is very similar to the exponential model, except that $|h|/a$ in the exponent now is in the power of 2. We fit this model in exactly the same way as we fit the exponential model, except we plot in Cartesian coordinates the quantities $\ln(-\tilde{\gamma}(h) + C_0 + C_1)$ versus $|h^2|$, for $h^2 < h_0^2$. The intercept for $|h|^2 = 0$ will give us $\ln C_1$ and the slope of the fitted line will give us $-\frac{1}{a^2}$, since $\ln(-\tilde{\gamma}(h) + C_0 + C_1) = -\frac{-|h|^2}{a^2} + \ln C_1$

The **linear model** is given by [1]

$$\tilde{\gamma}(h) = C_0 + C_1 \frac{h}{a} \text{ if } |h| > 0 \tag{9}$$

This model is the easiest to work with. We work just like in the case of the fractal model only that in that case we were plotting $\log \tilde{\gamma}(h)$ versus $\log h$ and here we simply plot $\tilde{\gamma}(h)$ versus h.

When a model is only locally applicable, we should not use in Kriging all available points in order to interpolate at a particular point. Instead, local Kriging should be used, where only points at distance $h \leq h_{max}$ should be used for the interpolation. For simplicity, we use a square window of size $(2h_{max} + 1) \times (2h_{max} + 1)$ around each point the value of which is to be estimated, rather than a circular window.



(a) Original Image (b) Subsampled image (c) RMSE = 158.19 (d) RMSE = 135.87

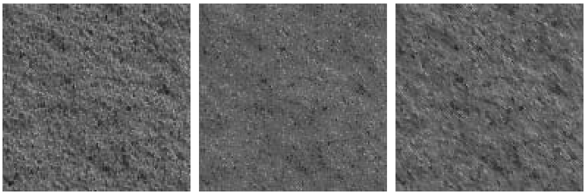
(e) RMSE = 101.72 (f) RMSE = 67.99 (g) RMSE = 44.21 (h) RMSE = 30.50

(i) RMSE = 23.73 (j) RMSE = 22.07

Fig. 5. Process of reconstruction for image IV from 4% of its pixels, using the exponential model.

4 Experiments

To experiment with interpolation methods we subsample the test images shown in Fig. 1, using sampling masks, with different numbers of pixels, uniformly distributed over the image. Then we see how well we can recover the values at the remaining positions, from the retained values. Fig. 2 shows an example of fitting the exponential model locally to the variograms constructed from the retained data of image III. There is some limitation when the interpolation is done by local Kriging when the number of points with known values inside the



Original image RMSE=13.82 RMSE=15.00

Fig. 6. The panel in the middle was reconstructed by global Kriging, while the panel on the right by local. Although the panel in the middle has a lower RMS error then the panel on the right, the panel on the right clearly looks as a better reconstruction.

Table 1. Root mean square error using the various models.

		Image				
Subsampling	Model	I	II	III	IV	V
4%	Fractal	N/C (N/C)	32 (33)	$L = G$ (23)	21 (20)	13 (N/C)
	Exponential	20 (20)	31 (30)	23 (24)	22 (21)	$L = G$ (25)
	Spherical	27 (21)	36 (30)	27 (24)	33 (21)	30 (25)
	Gaussian	51 (19)	33 (30)	$L = G$ (39)	72 (21)	30 (33)
	Linear	N/C (N/C)	33 (30)	25 (24)	22 (21)	25 (N/C)
6%	Fractal	N/C (N/C)	31 (31)	$L = G$ (21)	21 (19)	23 (N/C)
	Exponential	20 (20)	29 (28)	21 (21)	20 (19)	$L = G$ (23)
	Spherical	28 (N/C)	31 (28)	$L = G$ (21)	22 (19)	25 (23)
	Gaussian	55 (20)	33 (28)	$L = G$ (29)	23 (20)	29 (29)
	Linear	N/C (N/C)	32 (28)	26 (21)	23 (19)	23 (N/C)
9%	Fractal	18 (N/C)	18 (N/C)	$L = G$ (19)	19 (18)	21 (N/C)
	Exponential	18 (19)	28 (26)	19 (19)	20 (18)	$L = G$ (22)
	Spherical	24 (N/C)	30 (26)	$L = G$ (19)	22 (18)	24 (21)
	Gaussian	25 (20)	36 (26)	$L = G$ (29)	24 (19)	32 (26)
	Linear	N/C (19)	28 ((N/C)	24 (19)	20 (18)	21 (N/C)
26%	Fractal	16 (16)	23 (25)	$L = G$ (14)	19 (15)	17 (N/C)
	Exponential	17 (15)	23 (16)	13 (41)	18 (16)	$L = G$ (16)
	Spherical	19 (16)	24 (17)	$L = G$ (14)	18 (14)	$L = G$ (17)
	Gaussian	29 (22)	47 (25)	$L = G$ (18)	37 (20)	$L = G$ (25)
	Linear	17 (15)	23 (21)	19 (13)	18 (16)	33 (N/C)
45%	Fractal	15 (13)	20 (22)	$L = G$ (11)	15 (12)	13 (N/C)
	Exponential	15 (13)	21 (12)	10 (13)	16 (11)	$L = G$ (15)
	Spherical	16 (20)	21 (13)	$L = G$ (11)	16 (12)	$L = G$ (13)
	Gaussian	33 (21)	48 (42)	$L = G$ (33)	35 (28)	$L = G$ (32)
	Linear	15 (13)	21 ((N/C)	17 (11)	16 (11)	86 (N/C)

local window is too low. To overcome this problem, we only perform Kriging if the number of points with known values inside the local window is more then 5. We then perform subsequent iterations where the pattern gradually emerges by growing the "islands" of reconstructed pixels. Figures 3 – 5 show some examples of such growth. Table 1 summarises the reconstruction results by giving the root mean square error for each case. The error of the reconstruction using global Kriging is given inside paratheses. N/C in the table means Non-Convergence and $L = G$ means that the local fitting produced such a value of h_{max} that the "local" window was actually the whole image.

5 Conclusions

From the results presented here and many more results that we cannot report due to lack of space, we concluded that local Kriging produces visually better results than global Kriging. This is not confirmed by the RMS error values of

Table 1, in general, where global Kriging appears to create better reconstruction. Fig. 6 demonstrates this point. This shows that RMS is not the best way to assess the quality of a reconstruction texture of all models tried, as texture is a pattern rather than an exact deterministic image. The Gaussian model for variogram fitting performed the worst among all models tried.

Acknowledgments

This work was partly supported by a scholarship from the Commonwealth Commission UK, and partly by the Basic Technology grant GR/R87642/01 from the UK Research Councils.

References

1. Cressie, C. A. N.: Statistics for Spatial data. John Wiley and Sons, Inc (1993)Revised Edition
2. Davis, C. D.: Statistics and Data Analysis in Geology. John Wiley and Sons, Inc **5** (1973) 192–196
3. Gringarten, E., Deutsch, V. C.: Teacher's Aide Variogram Interpretation and Modeling. Mathematical Geology **33 (4)** (2001) 507–534
4. Isaaks, E. H., Srivastava, R. M.: An Introduction to Applied Geostatistics. New York; Oxford University Press (1989)
5. Kulatilake, W. S. H. P.,Um, J., Pan, G.: Requirements for Accurate Quantification of Self-Affine Roughness Using the Variogram Method. International Journal of Solids Structures **34 (31-32)** (1998) 4167–4172
6. Piroddi, R., Petrou, M.: Dealing with Irregular Samples (Book Chapter). Advances in Imaging and Electron Physics, P. W. Hawkes (Ed), Academic Press **132** (2004) 109–165

Spectral Methods in Image Segmentation: A Combined Approach

Fernando C. Monteiro[1,2] and Aurélio C. Campilho[1,3]

[1] INEB - Instituto de Engenharia Biomédica
[2] Escola Superior de Tecnologia e de Gestão de Bragança
Campus de Santa Apolónia, Apartado 134, 5301-857 Bragança, Portugal
monteiro@ipb.pt
[3] FEUP - Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
campilho@fe.up.pt

Abstract. Grouping and segmentation of images remains a challenging problem in computer vision. Recently, a number of authors have demonstrated a good performance on this task using spectral methods that are based on the eigensolution of a similarity matrix. In this paper, we implement a variation of the existing methods that combines aspects from several of the best-known eigenvector segmentation algorithms to produce a discrete optimal solution of the relaxed continuous eigensolution.

1 Introduction

The natural ability of the human visual system to separate an image into coherent segments or groups is extraordinary. This important phenomenon was studied extensively by the Gestalt psychologists, nearly a century ago [10]. They identified several key factors that contribute to human perceptual grouping process, including cues such as proximity, similarity, symmetry, continuity, common fate and familiarity.

An auspicious approach that has recently emerged uses spectral methods for image segmentation. These methods use the eigenvectors of a matrix representation of a graph to partition image into disjoint clusters with pixels in the same cluster having high similarity and points in different clusters having low similarity. A common characteristic among these techniques is the idea of clustering/separating pixels or other image elements using the dominant eigenvectors of a $n \times n$ matrix derived from the pair-wise affinities between pixels, where n denotes the number of pixels in the image. The affinity computed between pixels captures their degree of similarity as measured by one or more cues.

The general belief that these methods work is based on proofs that if segments are very dissimilar, spectral methods will be able to separate them [5]. In addition to, there is accumulated evidence that spectral methods find good or acceptable segmentation as judged by human on a variety of real data sets [3], i.e. these methods are effective in capturing perceptual organization features [2]. In spite of these facts, different authors still disagree on exactly which matrix

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 191–198, 2005.

and which eigenvectors they should use and how to proceed from the continuous eigenvectors to the discrete segmentation.

In section 3 we propose a new multiclass spectral algorithm that combines aspects from a set of algorithms to produce a discrete solution. The discretization is efficiently computed in an iterative way using singular value decomposition and non-maximum suppression. Mostly of previous works ([4], [5]) use a k-means clustering to get a discrete solution from eigenvectors. Although these methods can produce similar results to our approach, they may take twice as long to converge. Moreover, while for k-means a good initial estimation is crucial our method is robust to a random initialization.

2 Spectral Segmentation

2.1 Notation

We introduce some notation, before describing the algorithm in more detail. Let the symmetric matrix $W \in R^{n\times n}$ denote the weighted adjacency matrix for a graph $G = (V, E)$ with nodes V representing pixels and edges E whose weights capture the pair-wise affinities between pixels. Let A and B represent a bipartition of V, i.e., $A \cup B = V$ and $A \cap B = \emptyset$. The degree of dissimilarity between these two groups can be computed as total weight of the edges that must be removed to separate the groups. In graph theoretic language, it is called the *cut*:

$$cut\,(A, B) = \sum_{i\in A, j\in B} W\,(i, j) \ . \tag{1}$$

Although there are efficient computational algorithms to find partitions that minimizes the *cut* value, this criterion favours partitions which have small sizes [11]. Shi and Malik [8] presented an extension of the *cut* criterion, called *normalized cut* criterion:

$$ncut\,(A, B) = \frac{cut\,(A, B)}{links\,(A, V)} + \frac{cut\,(A, B)}{links\,(B, V)} \ , \tag{2}$$

where $links(A, V)$ is the total edges weights connecting nodes of A to all nodes in the graph, and $links(B, V)$ is similarly defined. This new criterion avoids the segmentation of separated nodes. If we define $links(A, A)$ as the total weights of edges connecting nodes within A, we can also define a measure for the degree of similarity within groups for a given partition. Using $links(A, V)$ as a normalization term, we can get *normalized links* such as:

$$nlinks\,(A, B) = \frac{links\,(A, A)}{links\,(A, V)} + \frac{links\,(B, B)}{links\,(B, V)} \ . \tag{3}$$

A simple calculation shows that $ncut\,(A, B) = 2 - nlinks\,(A, B)$. Hence minimizing the degree of dissimilarity between the groups and maximizing the degree of similarity within the group, can be satisfied simultaneously by the *normalized* *cut*. Therefore, this criterion favours both tight connections within partitions and loose connections between partitions. Among numerous partitioning criterion only *minimum cut* [11] and *normalized cut* have this duality property.

A common matrix representation of graphs is the Laplacian. Let D be the degree diagonal matrix of W such that $D_{ii} = \sum_j W_{ij}$, i.e. D_{ii} is the sum of the weights of the connections from node i to all other nodes in the graph W. Then the Laplacian of W is the matrix $L = (D - W)$.

2.2 Previous Works

We can classify spectral methods in two classes: recursive spectral segmentation [8] - these algorithms try to split the data into two partitions based on a single eigenvector and are then recursively used to generate more partitions; and multi-way spectral segmentation ([4], [5], [12]) - these algorithms use information from multiple eigenvectors to do a direct multi-way partition of data. Experimentally it has been observed that using more eigenvectors and directly computing a k way partitioning produces better results (e.g. [2], [4], [5]).

As we saw above, a good segmentation corresponds to a partitioning scheme that separates all the nodes of a graph by cutting off the weakest links among them, i.e. minimizes the *cut* value. Wu and Leahy [11] proposed a clustering method based on the minimum criterion that minimizes (1). However, as the authors also noted in their work, and since the *cut* increases with the number of edges going across the two clusters, the minimum cut criteria favours cutting small sets of isolated nodes in the graph.

Shi and Malik proposed to use a normalized similarity criterion to evaluate a partition. One key advantage of using the *normalized cut* is that it makes possible to find a good approximation to the optimal partition[1]. The approximation to the optimal partition can be found by computing:

$$min_x ncut\left(x\right) = \min_y \frac{y^T \left(D - W\right) y}{y^T D y} \quad , \tag{4}$$

subject to the constraints that $y\left(i\right) \in \{-1, 1\}$ and $y^T D\mathbf{1} = 0$. y is a binary indicator vector specifying the group identity for each pixel and $\mathbf{1}$ is the vector of all ones. Notice that the above expression is a Rayleigh quotient, so if we relax y to take on real values (instead of two discrete values), the minimization becomes equivalent to solving the generalized eigenvalue system,

$$D^{-1/2} \left(D - W\right) D^{-1/2} z = \mu z \quad , \tag{5}$$

where $z = D^{1/2} y$. Shi and Malik verified that for the two-class *normalized cut* criterion, the global optimum in the relaxed continuous domain is given by the second smallest generalized eigenvector. This eigenvector of W is thresholded in order to cut the image into two parts. This process can be continued recursively

[1] Minimizing *normalized cut* exactly is a NP-complete problem.

as desired. However, as Shi and Malik noted, there is no guarantee that the solution obtained will have any relationship to the correct discrete solution.

The Scott and Longuet-Higgins algorithm [7] constructs a matrix M whose columns are the first k eigenvectors of W, normalizes the rows of M and constructs a matrix $Q = MM^T$. It produces a good segmentation if Q has only 1's or 0's. They use a not normalized similarity matrix. In [9], Weiss proposed an interesting combination of the Shi and Malik algorithm and the Scott and Longuet-Higgins algorithm and proved that it produces the best result. Meila and Shi algorithm [4] uses a random walk view in terms of the stochastic matrix P, with elements P_{ij}, obtained by normalizing the rows of W to sum 1. $P = D^{-1}W$ or $P_{ij} = W_{ij}/D_i$. This matrix can be viewed as defining a Markov random walk over nodes V, with P_{ij} being the transition probability $p\,[i \rightarrow j|\,i]$.

Equation (5) can be solved by a simpler eigensystem:

$$Px = \lambda x \quad . \tag{6}$$

The eigenvalues of P are $1 = \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n \geq -1$ and the corresponding eigenvectors are $x_1, x_2, ..., x_n$. Then from (5) we get,

$$\mu_i = 1 - \lambda_i \text{ and } z_i = D^{1/2} x_i \quad . \tag{7}$$

for all $i = 1, ..., n$. Note that this ensures that the eigenvalues of P are always real and the eigenvectors are linearly independent. Meila and Shi [4] form a matrix X whose columns are the eigenvectors corresponding to the k largest eigenvalues of P and then cluster the rows of X as points in a k-dimensional space.

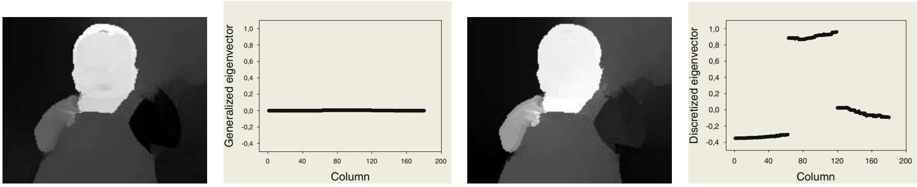
Ng *et al.* [5] use a different spectral mapping that behaves very similar to the Meila and Shi algorithm. It is proved that if the regions are well separated in the sense that the similarity matrix W is almost block diagonal, and if the sizes of the regions and the degrees of individual nodes don't vary too much, the rows of the X matrix cluster near k orthogonal vectors in R^k . This fact suggested the orthogonal initialization presented by Yu and Shi in [12].

3 Our Approach

We propose a multiclass algorithm based on a combined approach that uses random walk approach proposed by Meila and Shi [4] to create a normalized weight matrix P; Then, it solves an eigensystem and generates a matrix X, in the same manner as proposed by Ng *et al.* [5]; Finally, it uses a discretization process, proposed by Yu and Shi [12], more efficient than the k-means method, since it is robust to random initialization and converges faster.

3.1 The Algorithm

In an ideal case, the eigenvectors should only take on discrete values and the signs of the values can tell us exactly how to partition the graph. However, the eigenvectors can take on continuous values with very small variation among



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

Fig. 1. Continuous vs. discretized eigenvectors: **a.** A generalized continuous eigenvector of W. **b.** A horizontal cross section through the pixels in a. **c.** The discrete solution of the same eigenvector. **d.** A cross section through the pixels in c.

them. Figure 1 shows the relation between continuous and discretized eigenvectors. Although there is correct information in this continuous solution, it could be very hard to split the pixels into segments.

Our goal is to find the right orthogonal transform that leads to a discrete solution that satisfies the binary constraints of (4), yet it is closest to the continuous optimum. The result of such discrete solution is presented in Fig. 1.d. Note that pixels referring to the head are nearly all 1, while others are much smaller. From this result it is very easy to segment the image.

To obtain a discrete solution we follow the heuristics presented by Yu and Shi in [12]. Due to the orthogonal invariance of the eigenvectors, any continuous solution can be replaced by $\widetilde{Y}R$ for any orthogonal matrix $R \in \mathrm{R}^{k \times k}$. An optimal partition Y should satisfy the following conditions:

$$min\ \phi\left(Y,R\right) = \left\|Y - \widetilde{Y}R\right\|^2 \text{ with } Y \in \{0,1\}_{n \times k}\,,\ Y\mathbf{1}_k = \mathbf{1}_n,\ R^T R = I_k\ . \quad (8)$$

where $\mathbf{1}_k$ and $\mathbf{1}_n$ are vectors of all ones, and I_k is the identity matrix.

This can be solved by an iterative optimization process:

- Given R, we want to minimize $\phi\left(Y\right) = \left\|Y - \widetilde{Y}R\right\|^2$. The optimal solution is given by non-maximum suppression:

$$Y\left(i,m\right) = istrue\left(m = \arg\max\left[\widetilde{Y}\left(i,k\right)\right]\right),\ i \in V,\ m \in \{1..k\}\ . \quad (9)$$

We let the first cluster centroid be a randomly chosen row of the continuous solution $\widetilde{Y}$, and then repeatedly choose as the next centroid the row of $\widetilde{Y}$ that is closest to being 90^0 from all the centroids already picked.

- Given Y, we want to minimize $\phi\left(R\right) = \left\|Y - \widetilde{Y}R\right\|^2$. The solution is given by singular value decomposition (SVD):

$$U \cdot \varOmega \cdot \widetilde{U}^T = SVD\left(Y^T\widetilde{Y}\right)\ . \quad (10)$$

So, we can get,

$$R = \widetilde{U}U^T \text{ with } \min\phi\left(R\right) = 2\left(n - tr\left(\varOmega\right)\right)\ . \quad (11)$$

Such iterations monotonously decrease the distance between a discrete solution and the continuous optimum. The larger $tr\left(\varOmega\right)$ is, the closer Y is to $\widetilde{Y}R$.

Our segmentation algorithm consists of the following steps:

1. Set the diagonal elements W_{ii} to 0 and compute the normalized matrix P.
2. Let $1 = \lambda_1 \geq ... \geq \lambda_k$ be the k largest eigenvalues of P and $x_1, ..., x_k$ the corresponding eigenvectors. Form the matrix X by stacking the eigenvectors in columns.
3. Form the matrix $\widetilde{Y}$ from X by renormalizing each of X's rows to have unit length: $\widetilde{Y} = X \cdot Diag^{-1/2}\left(XX^T\right)$.
4. Initialize orthogonal matrix R with random lines of $\widetilde{Y}$.
5. Find the optimal discrete solution Y by (9).
6. Find the optimal orthogonal matrix R by (11).
7. While $|tr(\Omega) - \phi| > eps$ go to step 5.
8. Merge very similar neighbour regions which don't have edges among them.

3.2 Initialization of Affinity Matrix *W*

The quality of a segmentation based on the pair-wise similarities fundamentally depends on the weights that are provided as input. The weights should be large for pixels that should belong to the same group and small otherwise.

We associate to each pixel in the image a descriptor that captures brightness in a neighbourhood of the pixel. The similarity between two pixels is a function of the difference in their descriptors. Images are first convolved with oriented filter pairs (Fig. 2.b) to extract the magnitude of orientation energy (OE) of edge responses, as used by Malik *et al.* in [2]. At each pixel i, we can define the dominant orientation as $\theta^* = \arg\max OE_\theta$ and OE^* as the corresponding energy. The value OE^* is kept at the location of i only if it is greater than or equal to the neighbouring values. Otherwise it is replaced with a value of zero.

For each pair of pixels, pixel affinity is inversely correlated with the maximum contour energy encountered along the path connecting the pixels (Eq. 12). A large magnitude indicates the presence of an *intervening contour* and suggests that the pixels do not belong to the same segment.

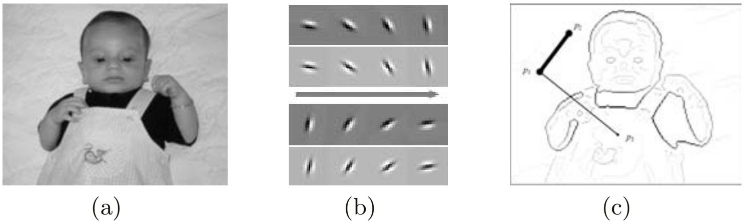
$$W(i,j) = \begin{cases} \exp\left[-\frac{\max_{t\in(0,1)} OE^*(s_i + t\cdot s_j)}{2\sigma_e^2 \cdot \max_l OE^*(s_l)}\right] & \text{if } \|s_i - s_j\| < r \\ 0 & \text{otherwise} \end{cases}, \tag{12}$$

where s_i denotes the spatial location of pixel i, l is the straight line between pixels, t is a binary value which takes value '1' if the phases of the pixels are different, and r defines the city-block distance.

Figure 2 illustrates the intuition behind this idea. The intensity values of pixels p_1, p_2 and p_3 are very similar. However, there is a contour among them, which suggests that p_1 and p_2 belong to one group while p_3 belongs to another.

3.3 Experiments

To test our algorithm, we applied it to a set of images from the Berkeley Segmentation Dataset [3]. It contains 12.000 manual segmentations of 1.000 images by 30 human subjects. Each image has been segmented by at least 5 subjects,

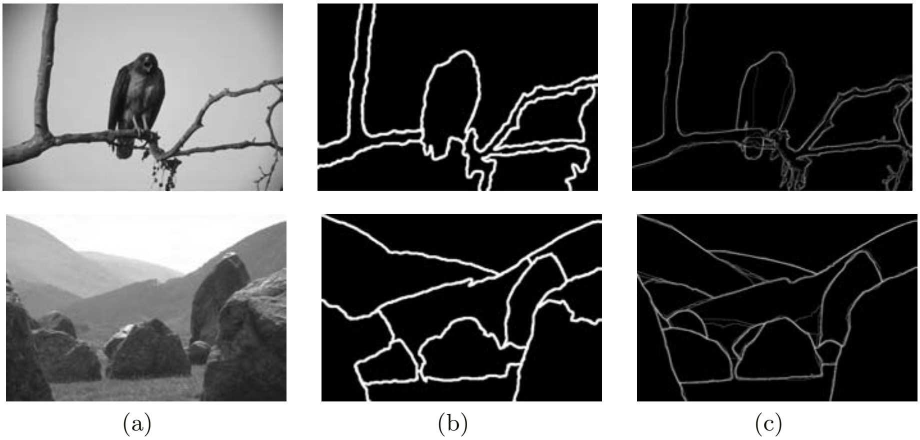


(a) (b) (c)

Fig. 2. Similarity matrix W is computed based on intensity edge magnitudes: **a.** The original image. **b.** The oriented filter pairs. **c.** Orientation energy.

so the ground truth is defined by a set of human segmentations. Martin *et al.* [3] declare two pixels to lie in the same segment only if all subjects declare that.

Figure 3 compares our results with the ground truth defined by Martin *et al.*. The column (Fig. 3.b) shows the results of some experiments with our algorithm and column (Fig. 3.c) represent the ground truth. Note that these ground truth images represent the probability that a segment will be chosen, if analysed by a person. We can see that our method reliably finds segments consistent with that an human would have chosen.



(a) (b) (c)

Fig. 3. Results of some experiments with the proposed algorithm: **a.** The original image. **b.** Our results. **c.** Berkeley ground truth.

3.4 Computation Time

The most time consuming part of the method is step 2, with a time complexity of $O\left(n^{3/2}k\right)$ using a Lanczos eigensolver [6]. The total time complexity of the algorithm is around $O\left(n^{3/2}k + 2nk^2\right)$. On a 1.4GHz Intel® Centrino™ processor, our method takes about 3 seconds on segmenting an 180 × 120 image, with $k = 10$, in C. This time could be greatly reduced by using the Nyström method proposed by Fowlkes *et al.* [1].

4 Conclusion

In this paper, we have presented a variation of the existing methods that combines aspects from different eigenvector segmentation algorithms. The heuristics are simple to implement as well as computationally efficient. Experimentally, we have demonstrated the potential of our approach for brightness and proximity image segmentation. However this model is general and can also be applied in a variety of image analysis. The improvement of the methodology can be achieved by designing better similarity distances between pixels. This can be done by using other cues such as texture or colour. However, good ways of combining these cues into one similarity matrix is still an open issue. Nevertheless, in the context of a specific application, dedicated similarity distances could be defined and lead to more precise segmentation results.

References

1. Fowlkes, C., Belongie, S., Chung, F., Malik, J.: Spectral Grouping Using Nyström Method, Trans. on Pattern Analysis and Machine Intelligence, 26(2), (2004) 214-225
2. Malik, J., Belongie, S., Leung, T., Shi, T.: Contour and Texture Analysis for Image Segmentation, International Journal of Computer Vision, 43(1), (2001) 7-27
3. 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, IEEE Int. Conf. on Computer Vision, Volume 2, (2001) 416-424
4. Meila, M., Shi, J.: Learning Segmentation by Random Walks, In Neural Information Processing Systems, (2000) 873-879
5. Ng, A., Jordan, M., Weiss, Y.: On Spectral Clustering: Analysis and an Algorithm, Advances in Neural Information Processing Systems 14, (2002) 849-856
6. Scott, D.S.: Solving Sparse Symmetric Generalized Eigenvalue without Factorization, SIAM Journal of Numerical Analisys, Volume 18, (1981) 102-110. Software available at: *http://www.netlib.org/laso/snlaso.f*
7. Scott, G.L., Longuet-Higgins, H.C.: Feature Grouping by "Relocalisation" of Eigenvectors of the Proximity Matrix, Proc. British Machine Vision Conf., (1990) 103-108
8. Shi, J., Malik, J.: Normalized Cuts and Image Segmentation, IEEE Trans. on Pattern Analysis and Machine Intelligence, 22(8), (2000) 888-905
9. Weiss, Y.: Segmentation Using Eigenvectors: A Unifying View, In International Conference on Computer Vision, (1999) 975-982
10. Wertheimer, M.: Laws of Organization in Perceptual Forms (partial translation), W. B. Ellis Editor, A Sourcebook of Gestalt Psychology, (1938) 71-88
11. Wu, Z., Leahy, R.: An Optimal Graph Theoretic Approach to Data Clustering: Theory and its Application to Image Segmentation, IEEE Trans. on Pattern Analysis and Machine Intelligence, Volume 15, (1993) 1101-1113
12. Yu, S., Shi, J.: Multiclass Spectral Clustering, International Conference on Computer Vision, Nice, France, (2003) 313-319

Mathematical Morphology in Polar-Logarithmic Coordinates. Application to Erythrocyte Shape Analysis

Miguel A. Luengo-Oroz[1], Jesús Angulo[1],
Georges Flandrin[2], and Jacques Klossa[3]

[1] Centre de Morphologie Mathématique, Ecole des Mines de Paris,
35 rue Saint-Honoré, 77305 Fontainebleau, France
[2] Unité de Télémédecine, Hôpital Universitaire Necker, 149 rue de Sèvres, 75743 Paris
[3] TRIBVN Company, Paris, France
{luengo,angulo}@cmm.ensmp.fr, Francegflandrin@wanadoo.fr,
jklossa@tribvn.com

Abstract. We present in this paper the application of mathematical morphology operators through a transformation of the Cartesian image into another geometric space, i.e. pol-log image. The conversion into polar-logarithmic coordinates as well as the derived cyclic morphology provides satisfying results in image analysis applied to round objects or spheroid-shaped 3D-object models. As an example of application, an algorithm for the shape analysis of the shape of red blood cells is given.

1 Introduction

A fundamental advantage of mathematical morphology [9] applied to image processing is that it is intuitive since it works directly on the spatial domain: the structuring elements considered as the "basic bricks" play the same role as frequencies do in the analysis of the sound. However, by using the discrete metrics and grids, which are more or less close to the Euclidian ones, we can not achieve the desired results when working on round objects.

It has been frequently suggested that the image should be transformed to other domains which would be adapted to the nature of the object or to the analysis that must be carried out (i.e. Fourier descriptors). This paper (extracted from [7]) proposes to use mathematical morphology operators through a transformation of the image into another geometric space that would present an intuitive nature. Therefore, we try to find a representation system which would present more advantages than the traditional Cartesian representation when processing and analizing images which contain some kind of radial symmetry, or in general, which have "a center". The selected transformation is the polar-logarithmic representation (or log-pol coordinates [12]). This mapping has already been used to map the visual cortex of primates [10](the photoreceptors of the retina are placed according to the same organization). Thus, this model of "log-pol fovea" is applied mainly to the artificial vision systems of robots [3], for which the need

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 199–206, 2005.

for real time processing of the visual information gives rise to the same problem regarding resource-optimization that the human visual system encounters. In addition, due to its scientific utility in describing fundamental aspects of human vision, the artificial "fovea" has been applied in order to assess the optical flow, to encode narrowband video, or else to recognize and track objects [2].

2 Log-Pol Coordinates

2.1 Definition

The polar-logarithmic transformation converts the original image (x, y) into another (ρ, ω) in wich the angular coordinates are placed on the vertical axis and the logarithm of the radius coordinates are placed on the horizontal one (furthermore a normalization has to be carried out in order to implement the transformation), see Fig. 1(a). More precisely, with respect to a central point (x_c, y_c): $\rho = \log(\sqrt{(x - x_c)^2 + (y - y_c)^2}$, $0 \leq \rho \leq \rho_{max}$; $\omega = \arctan \frac{(y-y_c)}{(x-x_c)}$, $0 \leq \omega < 2\pi$.

pseudocode direct transformation $(x, y) \Longrightarrow (\rho, \omega)$

```
for $\rho = 1 : R$ {; for $\omega = 1 : W$ {;
$x = \frac{\Delta X_{max}}{R} R^{\frac{\rho}{R}} \cos(\frac{2\pi\omega}{W}) + X_{central}$ ; $y = \frac{\Delta Y_{max}}{R} R^{\frac{\rho}{R}} \sin(\frac{2\pi\omega}{W}) + Y_{central}$
$ImageValue(\rho, \omega) = ImageValue(f(x, y))$ } }
```

2.2 Properties

Rotation. Because of the periodic nature of the angular component, rotations in the original Cartesian image become vertical cyclic shifts in the transformed log-pol image.

Scaling. The changes of size in the original image become horizontal shifts in the transformed image, according to the autosimilarity property of the exponential function, i.e. λ is the scale factor, $r' = \lambda r \Rightarrow \rho' = \log \lambda r = \log \lambda + \log r = cte + \rho$.

Choice of a Center. Due to the definition of the pol-log image, the choice of the center (x_c, y_c) is crucial. In fact, all the algorithms are sensitive to variations of the center, since the existence of a center is the principal prerequisite for all further developments. If the center point is not previously defined by the nature of the object, the choice of the center of gravity as central point is deemed adequate for most of cases.

If the goal is to analyze extrusions, the maximum of the distance function from the binary mask (the ultimate eroded set) can be considered as a satisfactory choice. This center maximizes the inscribed circumference within the object, however the main drawback is that this maximum can be multiple (set of regional maxima). In general, a better choice would correspond to the geodesic center defined as the minimum of the propagation function (if the set has no hole the propagation function reaches its minimum value at a unique point) [6].

3 Cyclic Morphology

3.1 Definition

Let $f(x, y)$ be an image defined on the discrete space $E \subset Z^2$, $(x, y) \in (Z \times Z)$, with values of the complete lattice $\mathcal{T}$ (for simplicity the complete lattice is considered to be $\mathcal{T} = Z$ or a subset from Z corresponding to the grey levels $\mathcal{T} = \{0, 1, \cdots, 255\}$). The extension of the operators from classical mathematical morphology to the log-pol representation is achieved by changing the support of the image in order to introduce the principle of periodicity. The log-pol transformation of the function $f(x, y)$ generates a new function image $\hat{f}(\rho, \omega) : E_{\rho,\omega} \longrightarrow \mathcal{T}$, where the support of the image is the space $E_{\rho,\omega}$, $(\rho, \omega) \in (Z \times Z_p)$ and where the angular variable $\omega \in Z_p$ is periodic with period p equivalent to 2π. A new relation of neighborhood is established where the points at the top of the image ($\omega = 0$) are neighbors to the ones at the bottom of the image ($\omega = p - 1$), therefore the edge connection should only take into account in the radial direction. The image can be seen as a strip where the superior and the inferior borders are joined, see Fig. 1(b).

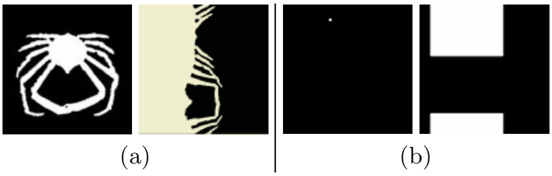


Fig. 1. (a) Example of conversion $(x, y) \rightarrow (\rho, \omega)$ ((x_c, y_c) corresponds to the body center of the crab). (b) Dilatation of one point by a square in log-pol coordinates.

The aim of this change is to preserve the invariance with respect to rotations in the Cartesian space, when morphological operations are done in the log-pol space, see an example in Fig. 2.

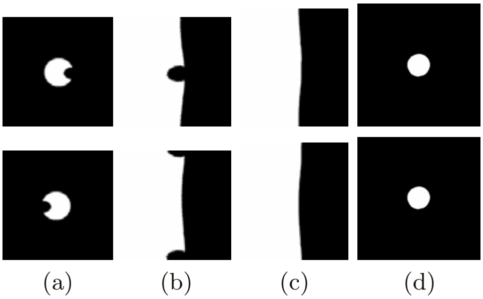


Fig. 2. (a) Original and 180^o rotation, (b) Direct transformation: pol-log, (c) Closing using as SE a centered square, (d) Invert transformation: Cartesian.

3.2 Implementation

In order to implement the new neighborhood relation and to be able to use morphological operators, two possibilities are considered:

- Modify the neighborhood relation and the basics operators code (erosion, dilation, etc.) by adding the operator "module of the size of the image in the direction of the cyclic coordinate". So if (ρ, ω) corresponds to the (x, y) axes, "y" should be substituted by "$y\ mod(y_{max})$" for the whole code.
- Extend the image along its angular direction by adding the top part of the image onto the bottom and the bottom part onto the top. The size of the vertical component from each part should be bigger than the size of the vertical component of the structuring element in order to avoid a possible edge effect. After having "cycled the image", morphological operators should be applied as usual and only the image corresponding to the initial mask should be kept. In Fig. 3 an illustrative example is shown. With this system all the existing code is recyclable.

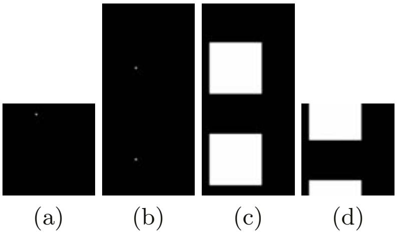


Fig. 3. Example of 2D Cyclic dilation: (a) Original , (b) "Cycled image", (c) Dilation by a square, (d) Original image mask: cyclic dilation.

3.3 Meaning of the Structuring Elements

The use of classic structuring elements (SE) in the log-pol image is equivalent to the use of "radial - angular" structuring elements in the original image. A vertical structuring element corresponds to an arc in the original image and a square corresponds to a circular sector (see fig. 4). For all the examples here presented, the center of the SE corresponds to the central point.

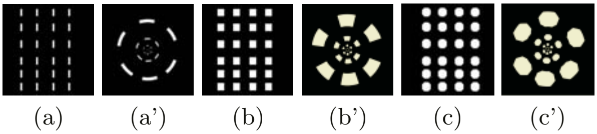


Fig. 4. Pol-log structuring elements (a,b,c) and their equivalence in the Cartesian space(a',b',c'). For all these exemples (x_c, y_c) is the central point of the image.

It is worth noting the fact that horizontal and vertical neighborhoods respectively acquire radial and angular sense in the original image; for instance, the transformation from a circumference is a vertical straight line.

4 Tools

Once cyclic morphology is defined, all the classic operators from mathematical morphology can be applied, giving at first view very interesting results for a certain kind of images. Some examples are given below.

4.1 Circular Filtering

One of the immediate applications is a method for extracting inclusions or extrusions from the contour of a relatively rounded shape with simple openings or closings [9]. The proportion of the vertical size of the structuring element with respect to the whole vertical size represents the angle affected in the original Cartesian image. With respect to a classical extraction in Cartesian coordinates, the choice of size is not as critical, making this a very advantageous point. It is due to the large zone plate in the openings/closings spectrum that is always found after a determined value (until the complete elimination of the object).

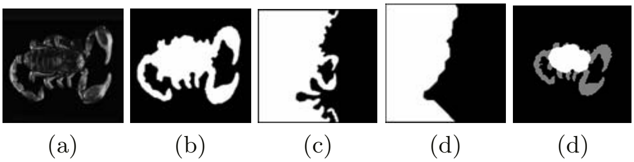


Fig. 5. Extremities segmentation from "Leiurius quinquestriatus": (a) Original, (b) Binary mask, (c) Log-pol transformation ((x_c, y_c) corresponds to the body center), (d) Opening with a vertical structuring element sized 20% of the whole image (i.e. 72^o), (e) Invert transformation.

4.2 Radial Skeleton

Let g be an image with only a connected component object. Let $\hat{g}$ be the log-pol transformation of g. If the chosen center to transformation $g \rightarrow \hat{g}$ is inside the object, the frontier of the object in $\hat{g}$ goes from the top of the image ($\omega = 0^o$) to the bottom ($\omega = 360^o$), and the connected component region resulting from the transformation of the object remains on the left of the edge ($\rho < \rho_{edge}$). If we apply an homotopic thinning [8] to $\hat{g}$ (according to the cyclic neighborhood); a skeleton is obtained mainly in the horizontal direction. This construction, when coming back to the Cartesian space, makes the skeleton to acquire a radial sense.

Therefore, we define a *radial inner skeleton* as the skeleton obtained by an homotopic thinning from the log-pol transformation of an objet. The invert transformation to Cartesian coordinates from the branches of the radial inner skeleton has radial sense and tends to converge on the center ($\rho = 0$). We also define the *radial outer skeleton* as the skeleton obtained by an homotopic thinning from the inverted image of the log-pol transformation of an object. The invert transformation to Cartesian coordinates from the branches of the radial outer skeleton has radial sense and this time, they tend to diverge to an hypothetical circumference in the infinity ($\rho \longrightarrow \infty$), see examples given in Fig. 6 and 8.

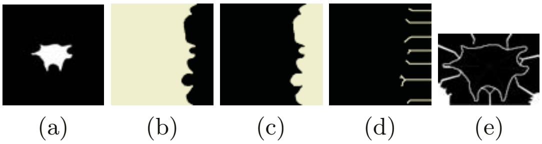


Fig. 6. Radial outer Skeleton: (a) Original , (b) Log-pol transformation (c) Inverted image , (d) Thinning, (e)Results in cartesian space.

5 Erythrocyte Shape Analysis: Inclusions and Extrusions Extraction Algorithm

In hematology, the visual analysis of the morphology of erythrocytes (size, shape, color, center,...) is of vital importance as it is well known that anomalies and variations from the typical red blood cell are associated with anemia or other illnesses [4]. In Fig. 7 a selection of abnormal erythrocytes is shown. We present hereafter one of the algorithms dedicated to the shape analysis developed in the MATCHCELL2 project [7], [1]. The aim of this algorithm is to extract the inclusions or extrusions from the erythrocyte shape, which is ideally round.

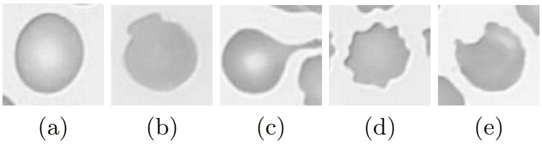


Fig. 7. (a) Normal erythrocyte, (b) "Mushroom" erythrocyte, (c) "Spicule" erythrocyte , (d) "Echinocyte" erythrocyte, (e) "Bitten" erythrocyte.

5.1 Algorithm

An algorithm for the extraction of extrusions and the identification of "mushroom" class is presented. It starts with the binary mask of the segmented erythrocyte, image (A), and the results correspond to image (G), see Fig. 8). If (G)$\neq \emptyset$ and the verifications are confirmed, it is classified as "mushroom" erythrocyte (we have considered the gravity center as the center of the log-pol transformation).

1°/ Log-pol transformation from (A) $\Rightarrow$ (B). 2°/ Radial inner skeleton from (B) $\Rightarrow$ (C). 3°/ Circular filtering: Residue from a vertical opening of 120^o (maximal admissible angle for the extrusion) from (B) $\Rightarrow$ (D)=extrusion candidates. 4°/ Geodesic reconstruction from (D) using as markers (D)$\cap$(C) $\Rightarrow$ (E). 5°/ Biggest connected set from (E) $\Rightarrow$ (F). We verify that $[Surface(F) > \mu_1 Surface(E)]$. 6°/ We verify that two branches of (D) reconstruct (F). 7°/ We verify that $[Surface(F)/sizeimage > \mu_2]$. 8°/Invert transformation from (F) to Cartesian coordinates $\Rightarrow$ (G)[1].

[1] μ_1 and μ_2 are fixed experimentally to 0.75 and 0.005.

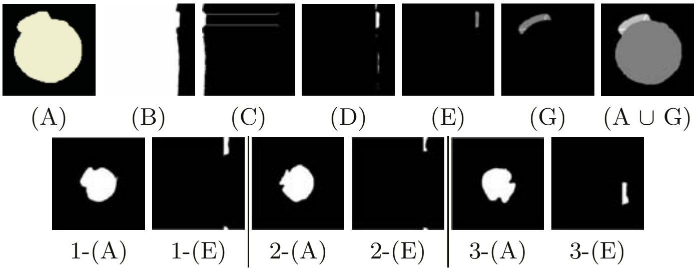


Fig. 8. Upper row, extrusion extraction algorithm for "mushroom" erythrocytes. Lower row, other examples of "mushroom" extraction (1,2,3).

Moreover, an analogous algorithm in order to extract the inclusions has been developed by applying the *radial outer skeleton* (step 2), and a closing instead of an opening (step 3).

5.2 Validation of the Approach

The algorithm has an efficient and robust performance in the extraction of inclusions and extrusions. The use of the skeleton in order to sieve the candidates gives much greater robustness than would a mere opening or closing. This procedure refines small connected components preselected as deformations. Furthermore, the examples corresponding to "bitten", "spicules" and "mushroom" have been correctly classified (see more examples and details in [7]).

6 Conclusions and Perspectives

The fundamental idea here presented is that the conversion of the image into another intuitive geometric representation can provide advantages over the traditional Cartesian representation. Regarding mathematical morphology, the key issue is to obtain structuring elements that are adapted to the nature of the objects to be analyzed, not by deforming them, but by transforming the image itself. The conversion into polar-logarithmic coordinates as well as the derived cyclic morphology appears to be a field that may provide satisfying results in image analysis applied to round objects or spheroid-shaped 3D-object models [7]. Basically, we have presented binary image processing and some of its basic tools, however the study of more complex morphological operators still remains, as well as a deeper developing of image processing for grey-scale or color.

Acknowledgements

We gratefully thank the "Groupe Français d'Hématologie Cellulaire" for supporting this study.

References

1. Angulo, J. (2003) *Morphologie mathématique et indexation d'images couleur. Application à la microscopie en biomédecine.* Ph.D. Thesis, Ecole des Mines de Paris.
2. Bernardino, A. and Santos-Victor, J. and Sandini, G.(2002)*Model Based Attention Fixation Using Log-Polar Images*, "Visual Attention Mechanisms", Plenum Press
3. Bolduc, M. and Levine, M. (1998) *A review of biologically motivated space-variant data reduction models for robotic vision*, CVIU, 69(2): 170−184.
4. Bronkorsta, P. and Reinders, M. (2000) *On-line detection of red blood cell shape using deformable templates*, Pattern Recognition Letters 21,413-424
5. Klossa J., G. Flandrin et J. Hemet. (2002) *Teleslide: better slide representativeness for digital diagnostic microscopy applications.*
6. Lantuejoul, C. et Maisonneuve, F. *Geodesic methods in quantitative image analysis.* Pattern Recognition, 17(2) : 177–187, 1984.
7. Luengo Oroz, M.A. (2004) *Morphologie mathématique en coordonnées logarithmique-polaires et méthodes géométriques de classification. Application à l'étude des érythocytes.* Master Thesis. Ecole des Mines de Paris.
8. Meyer, F.(1989) Skeletons and perceptual graphs. *Signal Processing,* 16 : 335–363.
9. Serra, J. (1982, 1988) *Image Analysis and Mathematical Morphology. Vol I & II*, Academic Press, London.
10. Schwartz, E. (1977) *Spatial mapping in the primate sensory projection : Analytic structure and relevance to perception*, Biological Cybernetics, 25:181−194.
11. Vincent, L.(1992) *Morphological area openings and closings for grayscale images.* In Proc. NATO Shape in Picture Workshop, Driebergen,Springer-Verlag.
12. Weiman, C. and Chaikin, G. (1979) *Logarithmic spiral grids for image processing and display.* Comp Graphics and Image Proc, 11:197−226.

Signal Subspace Identification in Hyperspectral Linear Mixtures*

José M.P. Nascimento[1] and José M.B. Dias[2]

[1] Instituto Superior de Engenharia de Lisboa and Instituto de Telecomunicações,
R. Conselheiro Emídio Navarro N 1, edifício DEETC, 1950-062 Lisboa, Portugal
Tel.:+351.21.8317237, Fax:+351.21.8317114
zen@isel.pt

[2] Instituto Superior Técnico and Instituto de Telecomunicações,
Av. Rovisco Pais, Torre Norte, Piso 10, 1049-001 Lisboa, Portugal
Tel.: +351.21.8418466, Fax: +351.21.841472
bioucas@lx.it.pt

Abstract. Hyperspectral applications in remote sensing are often focused on determining the so-called spectral signatures, i.e., the reflectances of materials present in the scene (endmembers) and the corresponding abundance fractions at each pixel in a spatial area of interest. The determination of the number of endmembers in a scene without any prior knowledge is crucial to the success of hyperspectral image analysis. This paper proposes a new mean squared error approach to determine the signal subspace in hyperspectral imagery. The method first estimates the signal and noise correlations matrices, then it selects the subset of eigenvalues that best represents the signal subspace in the least square sense.

1 Introduction

Hyperspectral remote sensing imagery is an important technology for monitoring the environment. Hyperspectral imagery is widely used in many applications such as land cover classification, mineral mapping, and detection of targets activities [1].

Hyperspectral systems have improved significantly through recent advances in sensor technology, being able to acquire many narrow contiguous bands of high spectral resolution in optical and infrared spectra [2, 3]. Hyperspectral sensors provide more detailed and accurate information of the spatial region than their multispectral ancestors, leading, however, to higher dimensional data sets.

Each pixel of an hyperspectral image can be considered as a vector in the space $\Re^L$, where L is the number of bands. Under the linear mixing scenario, the spectral vectors are a linear combination of a few vectors, the so-called endmember signatures. Therefore, the dimensionality of data is usually much lower

* This work was supported by the FCT, under the projects POSI/34071/CPS/2000 and PDCTE/CPS/49967/2003 and by DEETC of ISEL.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 207–214, 2005.

than the number of bands. A key problem in dimensionality reduction in hyperspectral imagery is the determination of the number of endmembers, termed intrinsic dimension (ID) of the data set. The estimation of the ID allows a correct dimension reduction and thus gains in computational time and complexity. Moreover, the projection of spectral vectors onto a subspace of lower dimension improves the signal-to-noise ratio (SNR).

There are basically two approaches for estimating ID [4]: global and local. The first estimates ID of data set as a whole. The second estimates ID using information contained in sample neighborhoods. The latter approach avoids the projection of data onto a lower-dimensional space. Projection techniques, which are generally used as global approaches, seek for the best subspace to project data by minimizing an objective function. For example principal component analysis (PCA) [5] seeks the projection that best represents data in the least square sense; maximum noise fraction (MNF)[6] or noise adjusted principal components (NAPC)[7] seeks the projection that optimizes the ratio of noise power to signal power. This is in contrast with PCA where no noise model is used.

Topological methods are local approaches that estimate the topological dimension of a data set [8]. For example curvilinear component analysis (CCA) [9] and curvilinear distance analysis (CDA) [10] are non-linear projections that are based on the preservation of the local topology.

Recently Harsanyi, Farrand, and Chang developed a Neyman-Pearson detection theory-based thresholding method, referred as HFC, to determine the number of spectral endmembers in hyperspectral data (see [11] chapter 17).

This paper proposes a method to estimate the number of endmembers in hyperspectral linear mixtures. The method first estimates the noise correlation matrix based on multiple regression theory, assuming spectral smoothness. Then an eigen-decomposition of the signal correlation matrix estimates is done.

To determine the signal subspace dimension, we identify the subset of eigenvalues that best represents, in the least square sense, the mean value of data set [12]. Since hyperspectral mixtures have nonnegative components, the projection of the mean value on any signal subspace eigenvector is always non-zero.

The paper is structured as follows. Section 2 describes the fundamentals of the proposed method. Section 3 evaluate the proposed algorithm using simulated and real data. Section 4 ends the paper by presenting some concluding remarks.

2 Subspace Estimation

Let $\mathbf{Y} = \left[\mathbf{Y}_1, \mathbf{Y}_2 \ldots \mathbf{Y}_N\right]$ be a $L \times N$ matrix of spectral vectors, one per pixel, where N is the number of pixels and L the number of bands. Assuming a linear mixing scenario, each observed spectral vector is given by

$$\begin{aligned} \mathbf{y} &= \mathbf{x} + \mathbf{n} \\ &= \mathbf{M}\mathbf{s} + \mathbf{n}, \end{aligned} \tag{1}$$

where $\mathbf{y}$ is an L-vector, $\mathbf{M} \equiv [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_p]$ is the mixing matrix ($\mathbf{m}_i$ denotes the ith endmember signature and p is the number of endmembers present in the covered area), $\mathbf{s} = [s_1, s_2, \ldots, s_p]^T$ is the abundance vector containing the fractions of each endmember (the notation $(\cdot)^T$ stands for vector transposed) and $\mathbf{n}$ models system additive noise.

Owing to physical constraints [13], abundance fractions are non-negative ($\mathbf{s} \succeq 0$) and satisfy the so-called positivity constraint $\mathbf{1}^T\mathbf{s} = 1$, where $\mathbf{1}$ is a $p \times 1$ vector of ones.

The correlation matrix of vector $\mathbf{y}$ is $\mathbf{R}_y = \mathbf{R}_x + \mathbf{R}_n$, where $\mathbf{R}_x = \mathbf{M}\mathbf{R}_s\mathbf{M}$ is the signal correlation matrix, $\mathbf{R}_n$ is the noise correlation matrix and $\mathbf{R}_s$ is the abundance correlation matrix. An estimate of the signal correlation matrix is given by

$$\widehat{\mathbf{R}}_x = \widehat{\mathbf{R}}_y - \widehat{\mathbf{R}}_n, \tag{2}$$

where $\widehat{\mathbf{R}}_y = \mathbf{Y}\mathbf{Y}^T/N$ is the sample correlation matrix of $\mathbf{Y}$, and $\widehat{\mathbf{R}}_n$ is an estimate of noise covariance.

Assuming that the spectral reflectance of endmembers varies smoothly, the noise correlation matrix $\widehat{\mathbf{R}}_n$ can be inferred based on multiple regression theory [14]. This consists in assuming that

$$\mathbf{Y}_i = \boldsymbol{\theta}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i, \tag{3}$$

where $\boldsymbol{\theta}_i = [\mathbf{Y}_1, \ldots, \mathbf{Y}_{i-1}, \mathbf{Y}_{i+1}, \ldots, \mathbf{Y}_L]$ is the explanatory data matrix, $\boldsymbol{\beta}_i = [\beta_1, \beta_2, \ldots, \beta_L]^T$ are the regression parameters, and $\boldsymbol{\epsilon}_i$ random errors. Vector $\boldsymbol{\beta}_i$ is inferred from $\mathbf{Y}$ for $i = 1, 2, \ldots, L$ by multiple regression theory. Finally, we compute $\widehat{\boldsymbol{\epsilon}}_i = \mathbf{Y}_i - \boldsymbol{\theta}_i\widehat{\boldsymbol{\beta}}_i$.

Figure 1 left, illustrates, based on simulation, the reflectance $\mathbf{x}$ and $\mathbf{x} + \mathbf{n}$ for a given pixel. Figure 1 right, presents true and estimated noise for the same pixel. Notice the similarity.

Matrix $\widehat{\mathbf{R}}_n$ is the sample covariance of the estimated noise $\widehat{\boldsymbol{\epsilon}}_i$.

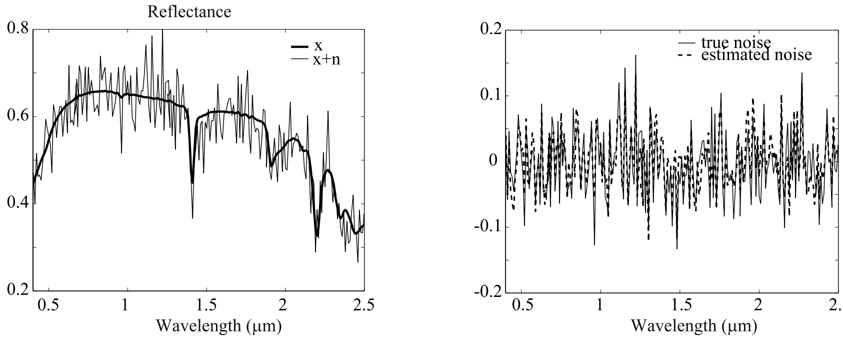


Fig. 1. Left: Illustration of the noise estimation based on spectral smoothness; Bold line: Reflectance of a pixel; Narrow line: Noise corrupted reflectance; Right: solid line: true noise; dashed line: estimated noise.

Let the singular value decomposition (SVD) of $\widehat{\mathbf{R}}_x$ be,

$$\widehat{\mathbf{R}}_x = \mathbf{E}\boldsymbol{\Sigma}\mathbf{E}^T, \tag{4}$$

where $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_k, \mathbf{e}_{k+1} \ldots, \mathbf{e}_L]$ is a matrix with the singular vectors ordered by the descendent magnitude of the singular values. The space $\Re^L$ can be splitted into two orthogonal subspaces: $< E_k >$ spanned by $\mathbf{E}_k = [\mathbf{e}_1, \ldots, \mathbf{e}_k]$ and $< E_k^\perp >$ spanned by $\mathbf{E}_k^\perp = [\mathbf{e}_{k+1}, \ldots, \mathbf{e}_L]$, where k is the order of the signal subspace.

Since hyperspectral mixtures have nonnegative components, the projection of the mean value of $\mathbf{Y}$ onto any eigenvector $\mathbf{e}_i$, $1 \leq i \leq k$, is always nonzero. Therefore, the signal subspace can be identified by finding the subset of eigenvalues that best represents, in the least square sense, the mean value of data set.

The sample mean value of $\mathbf{Y}$ is

$$\begin{aligned}\overline{\mathbf{y}} &= \frac{1}{N}\sum_{i=1}^{N} \mathbf{Y}_i \\ &= \mathbf{M}\sum_{i=1}^{N} \mathbf{s}_i + \frac{1}{N}\sum_{i=1}^{N} \mathbf{n}_i \\ &= \mathbf{c} + \mathbf{w},\end{aligned} \tag{5}$$

where $\mathbf{c}$ is in the signal subspace and $\mathbf{w} \sim \mathcal{N}(0, \mathbf{R}_n/N)$ [the notation $\mathcal{N}(\mu, \mathbf{C})$ stands for normal density function with mean μ and covariance $\mathbf{C}$]. Let $\mathbf{c}_k$ be the projection of $\mathbf{c}$ onto $< E_k >$. The estimation of $\mathbf{c}_k$ can be obtained by projecting $\overline{\mathbf{y}}$ onto the signal subspace $< E_k >$, i.e., $\widehat{\mathbf{c}}_k = \mathbf{P}_k\overline{\mathbf{y}}$, where $\mathbf{P}_k = \mathbf{E}_k\mathbf{E}_k^T$ is the projection matrix.

The first and second order moments of the estimated error $\mathbf{c} - \widehat{\mathbf{c}}_k$ are

$$\begin{aligned}E\big[\mathbf{c} - \widehat{\mathbf{c}}_k\big] &= \mathbf{c} - E\big[\widehat{\mathbf{c}}_k\big] \\ &= \mathbf{c} - E\big[\mathbf{P}_k\overline{\mathbf{y}}\big] \\ &= \mathbf{c} - \mathbf{P}_k\mathbf{c} \\ &= \mathbf{c} - \mathbf{c}_k \\ &\equiv \mathbf{b}_k,\end{aligned} \tag{6}$$

$$E\big[(\mathbf{c} - \widehat{\mathbf{c}}_k)(\mathbf{c} - \widehat{\mathbf{c}}_k)^T\big] = \mathbf{b}_k\mathbf{b}_k^T + \mathbf{P}_k\mathbf{R}_n\mathbf{P}_k^T/N, \tag{7}$$

where the bias $\mathbf{b}_k = \mathbf{P}_k^\perp\mathbf{c}$ is the projection of $\mathbf{c}$ onto the space $< E_k^\perp >$. Therefore the density of the estimated error $\mathbf{c} - \widehat{\mathbf{c}}_k$ is $\mathcal{N}(\mathbf{b}_k, \mathbf{P}_k\mathbf{R}_n\mathbf{P}_k^T/N)$,

The mean squared error between $\mathbf{c}$ and $\widehat{\mathbf{c}}_k$ is

$$\begin{aligned}\mathrm{mse}(k) &= E\big[(\mathbf{c} - \widehat{\mathbf{c}}_k)^T(\mathbf{c} - \widehat{\mathbf{c}}_k)\big] \\ &= \mathrm{tr}\{E\big[(\mathbf{c} - \widehat{\mathbf{c}}_k)(\mathbf{c} - \widehat{\mathbf{c}}_k)^T\big]\} \\ &= \mathbf{b}_k^T\mathbf{b}_k + \mathrm{tr}(\mathbf{P}_k\mathbf{R}_n\mathbf{P}_k^T/N).\end{aligned} \tag{8}$$

where tr$(\cdot)$ denote the trace operator. Since we do not know the bias $\mathbf{b}_k$, an approximation of (Eq. 8) can be achieved by using the bias estimates $\widehat{\mathbf{b}}_k = \mathbf{P}_k^{\perp}\overline{\mathbf{y}}$. However, $E\big[\widehat{\mathbf{b}}_k\big] = \mathbf{b}_k$ and $E\big[\widehat{\mathbf{b}}_k^T\widehat{\mathbf{b}}_k\big] = \mathbf{b}_k^T\mathbf{b}_k + \text{tr}(\mathbf{P}_k^{\perp}\mathbf{R}_n\mathbf{P}_k^{\perp T}/N)$, i.e., an unbiased estimate of $\mathbf{b}_k^T\mathbf{b}_k$ is $\widehat{\mathbf{b}}_k^T\widehat{\mathbf{b}}_k - \text{tr}(\mathbf{P}_k^{\perp}\mathbf{R}_n\mathbf{P}_k^{\perp T}/N)$. The criteria for the signal subspace order determination is then

$$\begin{aligned}\widehat{k} &= \arg\min_k\big(\widehat{\mathbf{b}}_k^T\widehat{\mathbf{b}}_k + \text{tr}(\mathbf{P}_k\mathbf{R}_n\mathbf{P}_k^T/N) - \text{tr}(\mathbf{P}_k^{\perp}\mathbf{R}_n\mathbf{P}_k^{\perp T}/N)\big)\\ &= \arg\min_k\big(\mathbf{y}^T\mathbf{P}_k^{\perp T}\mathbf{P}_k^{\perp}\mathbf{y} + 2\text{tr}(\mathbf{P}_k\mathbf{R}_n/N) - \text{tr}(\mathbf{I}/N)\big)\\ &= \arg\min_k\big(\mathbf{y}^T\mathbf{P}_k^{\perp}\mathbf{y} + 2\text{tr}(\mathbf{P}_k\mathbf{R}_n/N)\big), \qquad (9)\end{aligned}$$

where $\mathbf{I}$ is the identity matrix.

3 Experiments

3.1 Computer Simulations

In this section we test the proposed method in simulated scenes. The spectral signatures are selected from the U.S. geological survey (USGS) digital spectral library [15]. Abundance fractions are generated according to a Dirichlet distribution given by

$$p(\alpha_1, \alpha_2, \ldots, \alpha_p) = \frac{\Gamma(\mu_1 + \mu_2 + \ldots + \mu_p)}{\Gamma(\mu_1)\Gamma(\mu_2)\ldots\Gamma(\mu_p)}\alpha_1^{\mu_1-1}\alpha_2^{\mu_2-1}\ldots\alpha_p^{\mu_p-1}, \qquad (10)$$

where $0 \le \alpha_i \le 1$, $\sum_{i=1}^p \alpha_i = 1$, $E[\alpha_i] = \mu_i/\sum_{k=1}^p \mu_k$ is the expected value of the ith endmember fraction, and $\Gamma(\cdot)$ denotes the Gamma function.

The results next presented are organized into two experiments: in the first experiment the method is evaluated with respect to the SNR and to the number of endmembers (p). We define SNR as $SNR \equiv 10\log_{10}\big(E\big[\mathbf{x}^T\mathbf{x}\big]/E\big[\mathbf{n}^T\mathbf{n}\big]\big)$. In the second experiment, the method is evaluated when some endmembers are present in a few pixels of the scene.

In the first experiment, the hyperspectral scene has 10^4 pixels and the numbers of endmembers varies from 3 to 15. The abundance fractions are Dirichlet distributed with mean value $\mu_i = 1/p$, for $i = 1, \ldots, p$.

Fig. 2 left shows the evolution of the mean squared error, i.e., $\mathbf{y}^T\mathbf{P}_k^{\perp}\mathbf{y} + 2\text{tr}(\mathbf{P}_k\mathbf{R}_n/N)$ as a function of the parameter k, when $SNR = 35\,\text{dB}$ and $p = 5$. This figure shows that the minimum of the mean squared error occurs when $k = 5$, which is equal to the number of endmembers present in the image.

Table 1 presents the signal subspace order estimate as function of the SNR and of p. In this table it is compared the proposed method and the virtual dimensionality (VD), recently proposed in [11]. The VD was estimated by the NWHFC based eigen-thresholding method using the Neyman-Pearson test with the false-alarm probability set to $P_f = 10^{-4}$. The proposed method finds the correct ID for SNR larger than 25 dB, and under estimates ID as the SNR

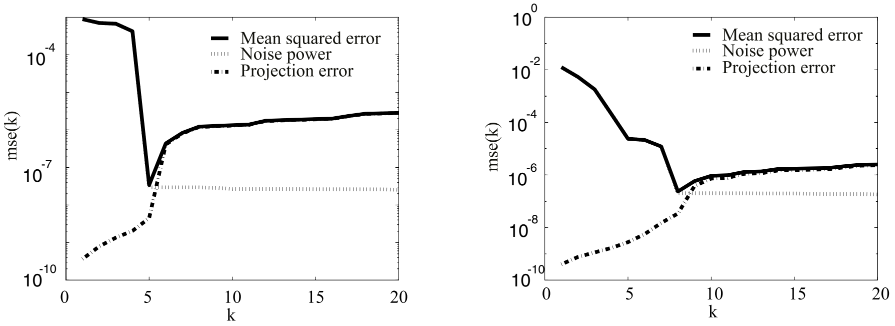


Fig. 2. Left: mean squared error versus k, with $SNR = 35\,\text{dB}$, $p = 5$; (first experiment) Right: mean squared error versus k, with $SNR = 35\,\text{dB}$, $p = 8$ (3 spectral vectors occur only on 4 pixels each; (second experiment).

Table 1. $\widehat{k}$ as function of SNR and of p; Bold: Proposed method; In brackets: VD estimation with NWHFC method and $P_f = 10^{-4}$.

$\widehat{k}$

Method	**New**	(VD)	**New**	(VD)	**New**	(VD)	**New**	(VD)	**New**	(VD)
SNR (in dB)	50		35		25		15		5	
$p = 3$	**3**	(3)	**3**	(3)	**3**	(4)	**3**	(4)	**3**	(2)
$p = 5$	**5**	(6)	**5**	(6)	**5**	(6)	**5**	(6)	**4**	(3)
$p = 10$	**10**	(11)	**10**	(11)	**10**	(9)	**8**	(8)	**6**	(2)
$p = 15$	**15**	(16)	**15**	(15)	**13**	(13)	**9**	(9)	**5**	(2)

decreases. In comparison with the NWHFC algorithm the proposed approach systematically yields better results.

In the second experiment $SNR = 35\,\text{dB}$ and $p = 8$. The first five endmembers have a Dirichlet distribution as in the previous experiment and the other three are forced to appear only in 4 pixels each one. Fig. 2 right show the mean squared error versus k, when $p = 8$. The minimum of mse(k) is achieved with $k = 8$. This means that the method is able to detect the rare endmembers in the image. However, this ability degrades as SNR decreases, as expected.

3.2 Cuprite Experiments

In this section, we apply the proposed method to real hyperspectral data collected by the AVIRIS [3] sensor over Cuprite, Nevada. Cuprite is a mining area in southern Nevada with mineral and little vegetation [16]. Cuprite test site, located approximately 200 Km northwest of Las Vegas is a relatively undisturbed acid-sulfate hidrothermal system near highway 95. The geology and alteration were previously mapped in detail [17, 18]. A geologic summary and a mineral map can be found in [16]. This site has been extensively used for remote sensing experiments over the past years [19, 20]. This study is based on subimage

(250 × 190 pixels and 224 bands) of a data set acquired on the AVIRIS flight 19 June 1997 (see Fig. 3 left). AVIRIS instrument covers the spectral region from $0.41\mu m$ to $2.45\mu m$ in 224 bands with $10nm$ bands. Flying at an altitude of $20km$, it has an instantaneous field of view (IFOV) of $20m$ and views a swath over $10km$ wide.

The method proposed when applied to this data set estimates $\widehat{k} = 23$ (see Fig. 3 right). According to the truth data presented in [16], there are 8 materials in these area. This difference is due to the following:

1. Rare pixels are not accounted in the truth data [16];
2. Spectral reflectances varies a little from pixel to pixel.

The bulk of spectral energy is explained with only a few eigenvectors. This can be observed from Fig. 3 center, where the accumulated signal energy is plotted as function of eigenvalues index. The energy contained in the first 8 eigenvalues is 99.94% of the total signal energy.

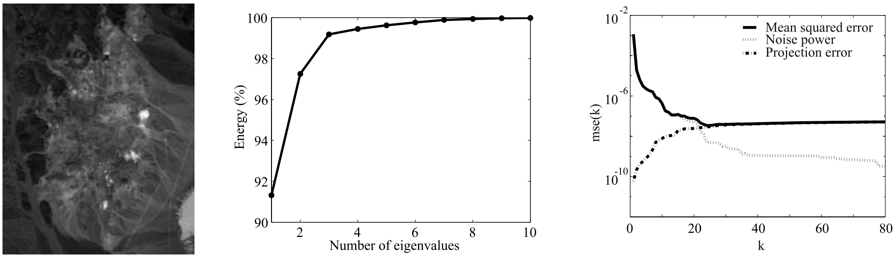


Fig. 3. Left: Band 30 (wavelength $\lambda = 667.3nm$) of the subimage of AVIRIS cuprite Nevada data set. Center: Percentage of signal energy as function of the number of eigenvalues; Right: mean squared error versus k for cuprite data set.

When VD was estimated by the HFC based eigen-thresholding method ($P_f = 10^{-3}$) on the same data set, this leads to an estimation of the number of endmembers equal to $\widehat{k} = 20$.

4 Conclusions

The determination of the signal subspace dimensionality is a difficult and challenging task. In this paper, we have proposed a method to estimate the dimensionality of hyperspectral linear mixtures. The method is based on the mean squared error criteria.

A set of experiments with simulated and real data leads to the conclusion that the method is an useful tool in hyperspectral data analysis yielding comparable or better results than the state-of-the-art methods.

References

1. Keshava, N., Mustard, J.: Spectral unmixing. IEEE Sig. Proc. Mag. **19** (2002) 44–57
2. Lillesand, T.M., Kiefer, R.W., Chipman, J.W.: Rem. Sens. and Image Interp. Fifth edn. John Wiley & Sons, Inc. (2004)
3. Vane, G., Green, R., Chrien, T., Enmark, H., Hansen, E., Porter, W.: The airborne visible/infrared imaging spectrometer (AVIRIS). Rem. Sens. of the Environ. **44** (1993) 127–143
4. Jain, A.K., Dubes, R.C.: Algorithms for clustering data. Prentice Hall, N. J. (1988)
5. Jolliffe, I.T.: Principal Component Analysis. Spriger Verlag, New York (1986)
6. Green, A., Berman, M., Switzer, P., Craig, M.D.: A transformation for ordering multispectral data in terms of image quality with implications for noise removal. IEEE Trans. Geosci. Rem. Sens. **26** (1994) 65–74
7. Lee, J.B., Woodyatt, S., Berman, M.: Enhancement of high spectral resolution remote-sensing data by noise-adjusted principal components transform. IEEE Trans. Geosci. Rem. Sens. **28** (1990) 295–304
8. Bruske, J., Sommer, G.: Intrinsic dimensionality estimation with optimaly topologic preserving maps. IEEE Trans. PAMI. **20** (1998) 572–575
9. Demartines, P., Hérault, J.: Curvilinear component analysis : A self-organizing neural network for nonlinear mapping of data sets. IEEE Trans. Neural Networks **8** (1997) 148–154
10. Lennon, M., Mercier, G., Mouchot, M., Hubert-Moy, L.: Curvilinear component analysis for nonlinear dimensionality reduction of hyperspectral images. In: Proc. of the SPIE. Volume 4541. (2001)
11. Chang, C.I.: Hyperspectral Imaging: Techniques for spectral detection and classification. Kluwer Academic, New York (2003)
12. Scharf, L.L.: Statistical Signal Processing, Detection Estimation and Time Series Analysis. Addison-Wesley Pub. Comp. (1991)
13. Manolakis, D., Siracusa, C., Shaw, G.: Hyperspectral subpixel target detection using linear mixing model. IEEE Trans. Geosci. Rem. Sens. **39** (2001) 1392–1409
14. Roger, R., Arnold, J.: Reliably estimating the noise in aviris hyperspectral imagers. International J. of Rem. Sens. **17** (1996) 1951–1962
15. Clark, R.N., Swayze, G.A., Gallagher, A., King, T.V., Calvin, W.M.: The u.s.g.s. digital spectral library: Version 1: 0.2 to 3.0 μm. Open file report 93-592, U.S.G.S. (1993)
16. Swayze, G., Clark, R., Sutley, S., Gallagher, A.: Ground-truthing aviris mineral mapping at cuprite, nevada,. In: Summaries of the Third Annual JPL Airborne Geosciences Workshop. Volume 1. (1992) 47–49
17. Ashley, R., Abrams, M.: Alteration mapping using multispectral images - cuprite mining district, esmeralda county,. Open file report 80-367, U.S.G.S. (1980)
18. Abrams, M., Ashley, R., Rowan, L., Goetz, A., Kahle, A.: Mapping of hydrothermal alteration in the cuprite mining district, nevada, using aircraft scanner images for the spectral region 0.46 to 2.36mm. Geology **5** (1977) 713–718
19. Goetz, A., Strivastava, V.: Mineralogical mapping in the cuprite mining district. In: Proc. of the Airborne Imaging Spectrometer Data Analysis Workshop, JPL Publication 85-41. (1985) 22–29
20. Kruse, F., Boardman, J., Huntington, J.: Comparison of airborne and satellite hyperspectral data for geologic mapping. In: Proc. of SPIE. Volume 4725. (2002) 128–139

Automatic Selection of Multiple Texture Feature Extraction Methods for Texture Pattern Classification*

Domènec Puig and Miguel Ángel Garcia

Intelligent Robotics and Computer Vision Group
Department of Computer Science and Mathematics, Rovira i Virgili University
Av. Països Catalans 26, 43007 Tarragona, Spain
{dpuig,magarcia}@etse.urv.es

Abstract. Texture-based pixel classification has been traditionally carried out by applying texture feature extraction methods that belong to a same family (e.g., Gabor filters). However, recent work has shown that such classification tasks can be significantly improved if multiple texture methods from different families are properly integrated. In this line, this paper proposes a new selection scheme that automatically determines a subset of those methods whose integration produces classification results similar to those obtained by integrating all the available methods but at a lower computational cost. Experiments with real complex images show that the proposed selection scheme achieves better results than well-known feature selection algorithms, and that the final classifier outperforms recognized texture classifiers.

1 Introduction

Texture classification consists of identifying the texture patterns present in an image given a set of known texture patterns (models) of interest (e.g., urban soil or crops in aerial images). In its most general form, the problem consists of identifying the texture pattern to which every image pixel belongs [11][3][10]. This problem differs from texture segmentation [4][6], which aims at finding regions of uniform texture within the image without identifying them.

In order to determine the texture pattern associated with the region in which a given pixel lies, it is first necessary to compute one or several *texture features* by evaluating some *texture feature extraction methods* (*texture methods* in short) in a neighborhood of the pixel. The features obtained after applying one or several texture methods can then be recognized by a pattern classifier.

A wide variety of texture methods have been proposed in the literature [11]. The majority of texture classifiers combine methods from the same family (e.g., Gabor filters) [9][11][12]. However, every family of texture methods is potentially useful for texture discrimination to a larger or lesser extent. Thus, it was shown in [3][10] that the proper integration of texture methods from different families leads to better classifica-

* This work has been partially supported by the Spanish Ministry of Science and Technology under project DPI2004-07993-C03-03.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 215–222, 2005.

tion results than those obtained with well-known texture classifiers based on methods from a single family.

However, it is not necessary to combine all the available texture methods for achieving good results. Frequently, after integrating several methods, further improvements are negligible compared to the increase in computation. In addition, it is also well known that classification rates may start degrading if too many texture methods are integrated due to the curse of dimensionality. Thus, determining a minimum subset of methods whose integration maximizes the final classification is still an open issue.

Feature selection for classification aims at selecting a feature subset without significantly decreasing the accuracy that the classifier reaches when it utilizes all the available features [2][5]. Hence, feature selection algorithms can be applied for obtaining a significant subset of texture methods given a specific texture discrimination problem. However, there is no consensus with respect to the utilization of those algorithms due to the amount of factors that affect their performance: the type and dimensionality of the data, the number of available features, the number of classes, the evaluation criterion utilized for determining the goodness of a subset of features, the search procedure for generating subsets of features, and the criterion used for stopping the search of new subsets.

In fact, the only search strategy that assures a unique optimal feature subset is the one based on an exhaustive search, by examining all the possible subsets of a desired size. However, this alternative is not feasible even if the dimension of the feature set is not excessively large, since the number of possible subsets increases combinatorially. Thus, a number of suboptimal selection techniques have been proposed [2][8] that apply trade-off solutions between computational efficiency and classification accuracy in order to generate subsets of features.

Although a large number of feature selection algorithms have been developed, they are still inefficient to use because they require users with knowledge about low-level details of the procedure [2][5][8]. From the previous surveys it follows that there is no single feature selection method that can be applied to all datasets or application fields. The choice of a feature selection method depends on various dataset characteristics: the ability to handle different data types, the ability to handle multiple classes, the ability to handle large datasets and the ability to handle noisy datasets.

This paper presents a new texture method selection scheme based on the texture classifier previously proposed in [3][10]. By taking into account the texture patterns to be classified, the proposed technique determines a reduced number of texture methods whose integration produces classification results comparable to or better than those obtained when all texture methods are integrated. Experimental results with complex textured images show that the proposed selection scheme leads to better classification results than when other well-known feature selection algorithms [2][8] are instead utilized. The final texture classifier outperforms recognized texture classifiers based on texture methods belonging to the same family.

The organization of this paper is as follows. Section 2 summarizes the texture classifier that is the basis of the proposed technique. Section 3 describes the proposed texture feature selection scheme. Section 4 shows experimental results of the integration of widely-used texture methods with the proposed technique, as well as a comparison with a well-known texture classification framework (MeasTex [12]). Conclusions and further improvements are finally presented in section 5.

2 Texture Classification

Let $\{\tau_1, \ldots, \tau_T\}$ be a set of T texture patterns of interest. Every texture τ_j is described by a set of sample images. Let **I** be an input textured image. In order to classify a target pixel $\mathbf{I}(x, y)$, a feature vector, $(\mu_1(\mathbf{I}(x, y)), \ldots, \mu_M(\mathbf{I}(x, y)))$, is determined. Each feature $\mu_i(\mathbf{I}(x, y))$ is obtained by applying a texture method μ_i to the pixels contained in a square window centered at $\mathbf{I}(x, y)$, whose size is experimentally set for each method. M texture methods are considered. The classification algorithm first introduced in [3] and later improved in [10] consists of four stages summarized below:

(a) *Supervised training stage.* A set of $M \times T$ likelihood functions $P_i(\mathbf{I}(x, y)|\tau_j)$ are defined based on the probability distributions P_{ij} corresponding to the evaluation of every method μ_i over the pixels of the sample images corresponding to each texture τ_j. Each distribution is defined in the interval $[MIN_{ij}, MAX_{ij}]$:

$$P_i(\mathbf{I}(x, y)|\tau_j) = P_{ij}(\mu_i(\mathbf{I}(x, y)) \in [MIN_{ij}, MAX_{ij}]) \tag{1}$$

(b) *Integration of multiple texture methods.* A linear opinion pool [1] combines the above likelihood functions:

$$P(\mathbf{I}(x, y)|\tau_j) = \sum_{i=1}^{M} w_{ij}\, P_i(\mathbf{I}(x, y)|\tau_j) \tag{2}$$

Every w_{ij} is computed as the average of the *KJ*-divergence between τ_j and the other models:

$$w_{ij} = d_{ij} / \sum_{r=1}^{M} d_{rj} \qquad d_{ij} = \frac{1}{T-1} \sum_{k=1, k \neq j}^{T} KJ_i(\tau_k, \tau_j) \tag{3}$$

The *Kullback J-divergence* [7], which measures the separability of two probability distributions is defined as:

$$KJ_i(\tau_a, \tau_b) = \int_0^1 (A - B)\log(A/B)\mathrm{d}u \tag{4}$$

with A and B being defined from the previous probability distributions: $A = P_{ia}(MAX_{ia}\, u + MIN_{ia}(1 - u))$ and $B = P_{ib}(MAX_{ib}\, u + MIN_{ib}(1 - u))$.

(c) *Maximum a Posteriori Estimation.* A set of T posterior probabilities are computed by applying the Bayes rule:

$$P(\tau_j|\mathbf{I}(x, y)) = \frac{P(\mathbf{I}(x, y)|\tau_j)P(\tau_j)}{\sum_{k=1}^{T} P(\mathbf{I}(x, y)|\tau_k)P(\tau_k)} \qquad P(\tau_j) = \sum_{i=1}^{M} w_{ij} / \sum_{i=1}^{M} \sum_{k=1}^{T} w_{ik} \tag{5}$$

$\mathbf{I}(x, y)$ is likely to belong to the texture class τ_j with the maximum posterior probability $P(\tau_j|\mathbf{I}(x, y))$.

(d) *Significance Test.* A significance level λ_j is defined based on two ratios utilized to characterize the performance of classifiers: sensitivity (S_n) and specificity (S_p) [10]. Pixel $\mathbf{I}(x, y)$ will be finally labelled as belonging to texture class τ_j iff $P(\tau_j|\mathbf{I}(x, y)) > \lambda_j$. Otherwise, it will be classified as unknown.

3 Automatic Selection of Texture Methods

The texture classifier summarized above is able to integrate any number of texture methods. This section presents an off-line selection algorithm that, given a set of texture patterns of interest and a set of texture methods, obtains a reduced subset of those methods whose integration allows to classify the given texture patterns similarly to or better than when all the texture methods are integrated.

Initially, the significance of every method is determined based on its performance in classifying the given patterns. All methods are then sorted in descending order of significance. Finally, a sequential forward generation procedure keeps adding new methods from the top of the sorted list until a performance criterion is maximized.

The *individual significance* S_{ij} of a texture method μ_i with respect to a texture pattern τ_j is estimated as follows. Let $\mathbf{I}_j$ be a sample image (or a collection of images) of texture τ_j. By applying (1) to $\mathbf{I}_j$ the likelihood $P_i(\mathbf{I}_j(x, y)|\tau_j)$ is obtained. Based on it, the Bayes rule is applied with all priors being $P_i(\tau_k) = 1/T$, $k \in [1, T]$:

$$P_i(\tau_j|\mathbf{I}_j(x, y)) = \frac{P_i(\mathbf{I}_j(x, y)|\tau_j)}{\sum_{k=1}^{T} P_i(\mathbf{I}_j(x, y)|\tau_k)} \tag{6}$$

Every pixel $\mathbf{I}_j(x, y)$ is then classified into the texture pattern that leads to the maximum posterior probability $P_i(\tau_j|\mathbf{I}_j(x, y))$. Let R_{ij} be the percentage of pixels from $\mathbf{I}_j$ that are correctly classified (classification rate) into texture τ_j when method μ_i is utilized. The individual significance S_{ij} is defined as the normalized classification rate:

$$S_{ij} = R_{ij} / \sum_{k=1}^{M} R_{kj} \qquad S_i = \sum_{j=1}^{T} S_{ij} \tag{7}$$

S_{ij} is defined in the interval [0, 1], with zero indicating that μ_i is unable to distinguish pattern τ_j, and one that μ_i is the only method able to classify that pattern. The global significance S_i of method μ_i for the given texture patterns is finally defined as the sum of individual significances S_{ij} associated with that method (7). The M texture methods μ_i are then sorted in descending order of S_i. This ordering is already useful in case the texture classifier utilizes a predefined number of methods lower than M.

However, the final goal consists of choosing a specific subset of methods out of the sorted list. Hence, the texture classifier from section 2 is subsequently applied to the sample images of the given texture patterns in order to compute average classification rates by integrating different subsets of methods, starting with the most significant

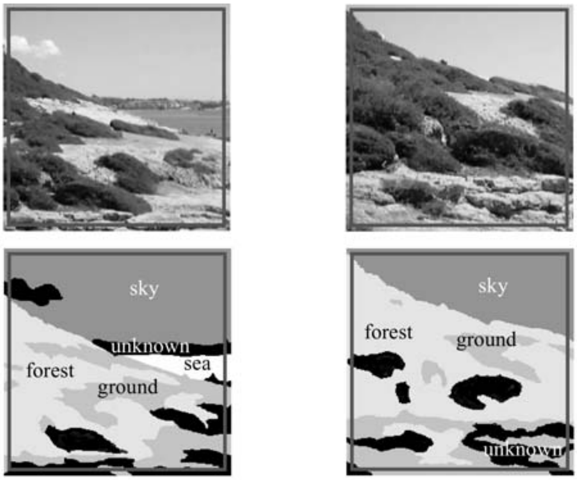


Fig. 1. Test images and ground-truth. Black regions in the latter correspond to unknown patterns. Dark squares enclose areas in which 17x17 pixel windows can be fitted

method and progressively adding a new method from the head of the sorted list until all methods are integrated. Let Φ and ϕ respectively be the maximum and minimum average classification rates obtained after the previous iterative process.

Let $R_{[1,m]}$ be the average classification rate obtained by integrating the first m methods, $m \in [1, M]$, from the head of the sorted list (m most significant methods). Both $R_{[1,m]}$ and m are normalized between zero and one: $\overline{R}_{[1,m]} = (R_{[1,m]} - \phi)/(\Phi - \phi)$ and $\overline{m} = (m-1)/(M-1)$. A performance measure ranging between zero and one is defined as:

$$\rho_{[1,m]} = \frac{1}{3}\left(2\frac{\overline{R}_{[1,m]}+1}{\overline{m}+1} - 1\right) \qquad (8)$$

which gets its maximum value in case the maximum classification rate Φ were obtained with the first, most significant method ($\Phi = R_1$), and its minimum value if the minimum rate ϕ were obtained when all the M available methods are integrated ($\phi = R_M$). In the end, the subset formed by the η most significant methods is chosen, with η being:

$$\eta = \arg\max_{m} R_{[1,m]}\rho_{[1,m]} \qquad (9)$$

4 Experimental Results

The proposed technique has been evaluated on real outdoor images containing complex textured surfaces. Fig. 1(*top row*) shows two of those input images. Four texture patterns of interest have been considered: “sky”, “forest”, “ground” and “sea”. A set of sample images representing each of those patterns was extracted from the image database and utilized as the training set for the classifiers.

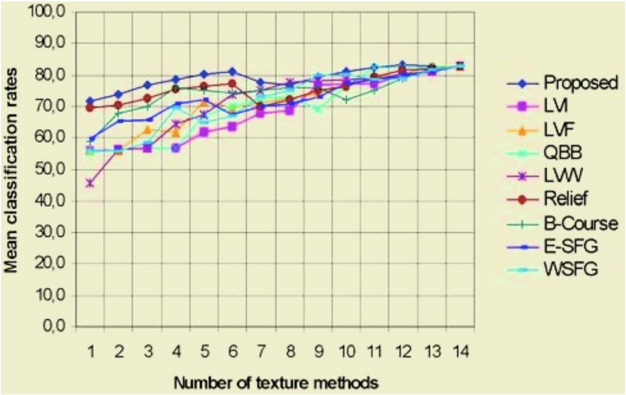


Fig. 2. Classification rates for the sample images of the 4 texture patterns by integrating a varying number *m* of texture methods. (Proposed) *m* methods sorted by significance. (Others) *m* methods selected with previous feature selectors

Fig. 1(*bottom*) shows the ground-truth classification of the test images by considering the texture patterns of interest. Black areas represent image regions that do not belong to any of the sought texture patterns —a supervised texture classifier aims at identifying a set of texture patterns in an input image, not at segmenting all of the image regions. Pixels that belong to those “unknown” texture patterns have not been taken into account in the classification rates presented below.

Considering previous surveys (e.g., [11]), 14 widely-used texture methods evaluated on 17x17 pixel windows have been considered to be integrated with the texture classifier summarized in section 2: four *Laws filter masks* (*R5R5, E5L5, E5E5, R5S5*), two *wavelet transforms* (*Daubechies-4, Haar*), four *Gabor filters* with different wavelengths (8, 4) and orientations (0º, 45º, 90º, 135º), three *statistics* (*variance*, *skewness*, *homogeneity*) and the *fractal dimension.* Based on the sample images of the 4 texture patterns of interest, the 14 methods were sorted in descending order of significance (7).

The texture classifier was then applied to the given sample images 14 times, initially considering only the most significant texture method and then adding a new method from the head of the sorted list at a time, until all 14 methods were integrated. Fig. 2(Proposed) shows the mean classification rates obtained for every subset of methods.

If (9) is applied, a trade-off solution consisting of the 5 most significant methods is obtained: *Daubechies-4*, *Haar*, *E5E5*, *variance*, *E5L5*. This combination leads to classification rates similar to those obtained with all 14 methods (see Table 1) but at a much lower cost (one third of methods).

The proposed selection scheme has been compared to 8 general purpose feature selection algorithms [2] (*LVI*, *LVF*, *QBB*, *LVW*, *Relief*, *B-Course*, *E-SFG*, *WSFG*). All pixels of the sample images were classified with the proposed texture classifier, which was executed 14 times for each feature selection algorithm, every time with a different number of methods from 1 to 14. The integrated methods were determined by the corresponding feature selector, given the desired number of methods and using as training

Table 1. Classification rates (%) for the images in Fig. 1 with the proposed classifier (integrating 14 and 5 selected methods) and with MeasTex

Texture Classifier	**Fig. 1 (*left*)**	**Fig. 1 (*right*)**
Proposed classifier with all 14 methods	75.5	73.0
Proposed classifier with 5 selected methods	75.0	72.2
MeasTex (12 Gabor filters, MVG)	72.3	65.7
MeasTex (12 Gabor filters, 5-NN)	73.1	68.2
MeasTex (4 Fractal features, MVG)	63.9	60.4
MeasTex (4 Fractal features, 5-NN)	64.5	63.6
MeasTex (4 GLCM statistics, MVG)	65.4	63.1
MeasTex (4 GLCM statistics, 5-NN)	51.7	56.5

features the outcome of all 14 texture methods applied to the same sample images. Fig. 2 shows the average classification rates corresponding to those experiments. In the majority of cases, the subsets obtained by applying the proposed significance-based ordering led to the best classification rates, including the subset with 5 methods.

The proposed texture classifier complemented with the new selection scheme has also been compared to the widely-used texture classifiers included in the MeasTex framework [12]. The same sample images utilized before were also used as the training dataset for MeasTex.

Table 1 gives the pixel classification rates corresponding to the test images from Fig. 1. The first two rows correspond to the application of the proposed texture classifier respectively fed with all the available methods and the 5 selected ones. Both S_n and S_p (see *significance test* in section 2) were set to 95%. The remaining rows show the results produced by MeasTex for different combinations of texture families {Gabor, Fractal, GLCM} and classification algorithms {MVG, 5-NN}. The proposed classifier produces the best results even when MeasTex integrates more methods.

Fig. 3(*top*) shows the classification maps for the test images in Fig. 1 after applying the texture classifier fed with 5 methods chosen according to the proposed selection scheme. Fig. 3(*bottom*) shows the best classification maps given by MeasTex according to Table 1: 12 Gabor filters, 5-NN. The other MeasTex results are much worse qualitatively.

5 Conclusions

This paper presents a new technique for selecting a subset of texture methods whose integration with a previously proposed texture classifier produces classification results similar to or better than the results obtained when all the available texture methods are integrated. Experimental results with complex real images show that the proposed selection scheme is more advantageous than general purpose feature selection algorithms, and that a texture classifier which properly integrates those methods outperforms widely-recognized texture classifiers based on texture methods from a

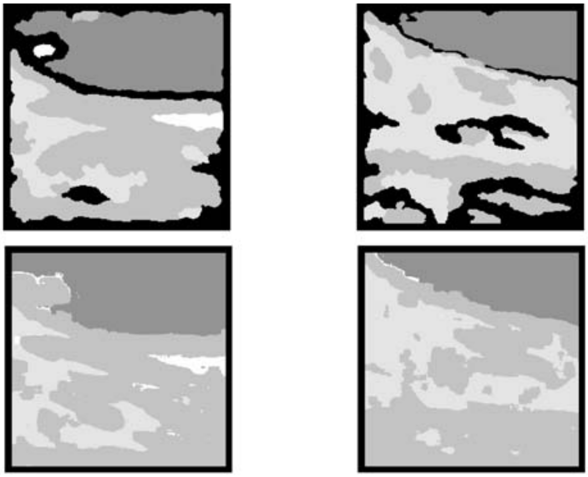


Fig. 3. (*top*) Classification maps with the proposed classifier by integrating the 5 selected methods. (*bottom*) Best classification maps with MeasTex (12 Gabor filters, 5-NN)

same family both quantitatively and qualitatively. Further work will consist of extending the proposed technique to unsupervised classification and segmentation, by automatically generating texture patterns from input images.

References

1. Berger, J.: Statistical Decision Theory and Bayesian Analysis. Springer (1985)
2. Dash, M., Liu, H.: Feature Selection for Classification. Intelligent Data Analysis. Elsevier (1997) 131-156
3. Garcia, M.A., Puig, D.: Improving Texture Pattern Recognition by Integration of Multiple Texture Feature Extraction Methods. 16th IAPR ICPR, Quebec, Canada (2002) 7-10
4. Hofmann, T., Puzicha, J., Buhmann, J.M.: Unsupervised Texture Segmentation in a Deterministic Annealing Framework. IEEE Trans. on PAMI, 29(8), (1998) 803-818
5. John, G.H., Kohavi, R., Pfleger, K.: Irrelevant Features and the Subset Selection Problem. 11th Int. Conf. on Machine Learning, New Brunswick NJ, USA (1994) 121-129
6. Malik, J., and others: Contour and Texture Analysis for Image Segmentation. In: Boyer, K.L., Sarkar, S. (eds.): Perceptual Organization for Artificial Vision Sys. Kluwer Ac. (2000)
7. Mathiassen, J.R., Skavhaug, A., Bo, K.: Texture Similarity Measure Using Kullback-Leibler Divergence between Gamma Distributions. 7th. ECCV, Denmark (2002) 133-147
8. Molina, L.C., Belanche, L., Nebot, A.: Feature Selection Algorithms: A Survey and Experimental Evaluation. Int. Conf. on Data Mining, Japan (2002) 306-313
9. Ojala, T., Pietikainen, M., Harwood, D.: A Comparative Study of Texture Measures with Classification Based on Feature Distributions. Pattern Recognition, 29(1), (1996) 51-59
10. Puig, D., Garcia, M.A.: Pixel Classification Through Divergence-Based Integration of Texture Methods with Conflict Resolution. IEEE ICIP, Barcelona, Spain (2003)
11. Randen, T., Husoy, J.H.: Filtering for Texture Classification: A Comparative Study. IEEE Trans. PAMI, 21(4), (1999) 291-310
12. Smith, G., Burns, I.: Measuring Texture Classification Algorithms. Pattern Recognition Letters 18, (1997) 1495-1501. (MeasTex Image Texture Database and Test Suite)

Dynamic Texture Recognition Using Normal Flow and Texture Regularity

Renaud Péteri[1] and Dmitry Chetverikov[2]

[1] Centre for Mathematics and Computer Science (CWI),
Kruislaan 413, 1098SJ Amsterdam, The Netherlands
Renaud.PETERI@mines-paris.org
[2] MTA SZTAKI - Hungarian Academy of Sciences
1111 Budapest, Kende u. 13-17, Hungary

Abstract. The processing, description and recognition of dynamic (time-varying) textures are new exciting areas of texture analysis. Many real-world textures are dynamic textures whose retrieval from a video database should be based on both dynamic and static features. In this article, a method for extracting features revealing fundamental properties of dynamic textures is presented. These features are based on the normal flow and on the texture regularity though the sequence. Their discriminative ability is then successfully demonstrated on a full classification process.

1 Introduction

1.1 Context

The amount of available digital images and videos for professional or private purposes is quickly growing. Extracting useful information from these data is a highly challenging problem and requires the design of efficient content-based retrieval algorithms. The current MPEG-7 standardization (also known as the "Multimedia Content Description Interface") aims at providing a set of content descriptors of multimedia data such as videos. Among them, texture [12] and motion [4] were identified as key features for video interpretation. Combining texture and motion leads to a certain type of motion pattern known as *Dynamic Textures* (DT). As the real world scenes include a lot of these motion patterns, such as trees or water, any advanced video retrieval system will need to be able to handle DT. Because of their unknown spatial and temporal extend, the recognition of DT is a new and highly challenging problem, compared to the static case where most textures are spatially well-segmented.

1.2 Dynamic Texture Recognition

Dynamic texture *recognition* in videos is a recent theme, but it has already led to several kinds of approaches. In *reconstructive approaches* ([11] or [10]), the recognition of DT is derived from a primary goal which is to identify the parameters of a statistical model 'behind' the DT. *Geometrical approaches* ([7], [13])

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 223–230, 2005.

consider the video sequence as a 3D volume (x, y and time t). Features related to the DT are extracted by geometric methods in this 3D space. *Qualitative motion recognition approaches* ([6], [1], [8]) are based on the human ability to recognize different types of motion, both of discrete objects and of DT. The aim is not to reconstruct the whole scene from motion, but to identify different kinds of DT.

1.3 Goals

In the context of video retrieval, we aim at supporting queries involving natural and artificial quasi-periodic DT, like fire, water flows, or escalator. In the following sections, a method for extracting features revealing the fundamental properties of DT is presented (section 2). These features are based on the normal flow and on the texture regularity though the sequence. Their discriminative ability is then tested in a full classification process (section 3).

2 Feature Extraction

2.1 The Normal Flow as a Medium of Motion Information

Definition. The quantitative approach to dynamic texture recognition is based on the assumption that computing full displacements is not necessary, is time consuming (because of regularization matters) and is moreover not always accurate. The partial flow measure given by the *normal* flow can provide a sufficient information for recognition purpose. By assuming constant intensity along 2D displacement, one can write the optical flow equation [5]:

$$\boldsymbol{v}(p).\boldsymbol{\nabla} I(p) + I_t(p) = 0 \tag{1}$$

with p the pixel where the optical flow is computed, $I_t(p)$ the temporal derivative of the image at p, and $\boldsymbol{\nabla} I(p)$ its gradient at p. Only the component of the optical flow parallel to $\boldsymbol{\nabla} I(p)$ can be deduced (known as the aperture problem). It then gives the *normal flow* $\boldsymbol{v_N}(p)$ ($\boldsymbol{n}$ is a unit vector in the direction of $\boldsymbol{\nabla} I(p)$):

$$\boldsymbol{v_N}(p) = -\frac{I_t(p)}{||\boldsymbol{\nabla} I(p)||}\,\boldsymbol{n} \tag{2}$$

Why Is the Normal Flow Suitable for Dynamic Textures? The normal flow field is fast to compute and can be directly estimated without any iterative scheme used by regularization methods [5]. Moreover, it contains both temporal and structural information on dynamic textures: temporal information is related to moving edges, while spatial information is linked to the edge gradient vectors. Its drawback is its sensitivity to noise, which can be reduced by smoothing or applying a threshold on spatial gradients.

Extraction of the Normal Flow Fields. In order to extract numerical features characterizing a DT, the normal flow fields are computed for each DT.

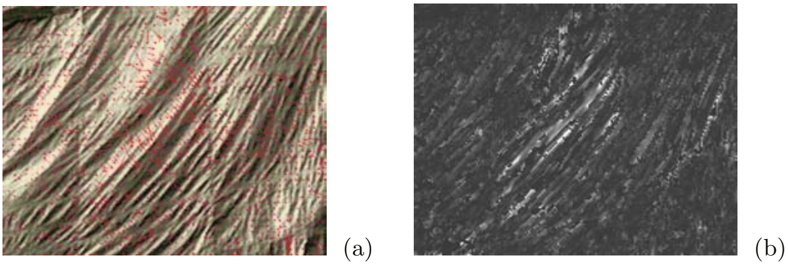


Fig. 1. Normal flow field (a) and its norm (b) on the 'plastic' sequence.

For reducing the noise-sensitivity of the normal flow, a Deriche [3] blurring filter, set with $\sigma = 1$, is applied to the sequence, followed by a linear histogram normalization. Image regions with low spatial gradients are masked through an automatic threshold on the spatial gradient, as values of their motion vectors would not be significant. The normal flow is then computed according to formula (2), excluding the 'masked' pixels.

Figure 1 illustrates the computation of the normal flow field (1a) and its norm (1b) on a waving 'plastic' sequence. The texture of the ripples created by the wind on the plastic sheet is well-visible.

2.2 Spatiotemporal Regularity Features

In this section, we summarize the regularity filtering method [2], then present our temporal regularity features used to classify dynamic textures.

The regularity method quantifies periodicity of a (static) texture by evaluating, in polar co-ordinates, the periodicity of its autocorrelation function. Consider a digital image $I(m, n)$ and a spacing vector (d_x, d_y). Denote by $\rho_{xy}(d_x, d_y)$ the normalized autocorrelation of $I(m, n)$. We obtain ρ_{xy} via the FFT using the well-known relation between the correlation function and the Fourier transform. The polar representation $\rho_{pol}(\alpha, d)$ is then computed on a polar grid (α_i, d_j) by interpolating $\rho_{xy}(d_x, d_y)$ in non-integer locations. The negated matrix is then used, referred to as the polar interaction map: $M_{pol}(i, j) = 1 - \rho_{pol}(i, j)$. (See figure 2.)

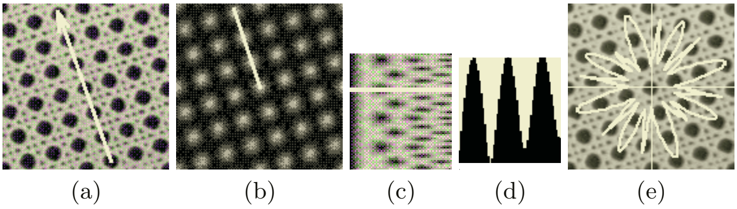


Fig. 2. (a) A pattern and a direction within the pattern. (b) Autocorrelation function. (c) Polar interaction map. (d) Contrast function for the direction. (e) Polar plot of $R(i)$ overlaid on the pattern.

A row of $M_{pol}(i,j)$ is called a contrast function. A periodic texture has contrast functions with deep and periodic minima. Our definition of regularity [2] quantifies this property. For each direction i, the algorithm [2] computes directional regularity $R(i) = R_{pos}(i)R_{int}(i)$, where $R_{pos}(i)$ describes the periodicity of the layout of the elements comprising the pattern, while $R_{int}(i)$ indicates how regular (stable) the intensity of the elements is. Based on $R(i)$, a number of texture regularity features can be computed. In this study, we only use the *maximal regularity* $M_R = \max_i R(i)$. $0 \leq M_R \leq 1$, with 0 indicating a random, 1 a highly regular pattern. $M_R \geq 0.25$ means visually perceivable periodicity.

The maximal regularity is computed for a set of overlapping windows covering the image. This procedure is called *regularity filtering*. The window size W is determined by the period of the structures to be detected: the filter responds to a structure if more than two periods are observed.

When applied to a dynamic texture, the method evaluates the temporal variation of the maximal regularity. For each frame t of a sequence, M_R is computed in a sliding window. Then the largest value is selected, corresponding to the most periodic patch within the frame. This provides a maximum periodicity value, $P(t)$, for each t.

2.3 Extracted Features

Six features have been defined for characterizing a dynamic texture. These features are based on the normal flow and on the texture regularity. (See table 1.)

Table 1. Features characterizing dynamic textures.

Normal flow		*Regularity*
1. Divergence	3. Peakiness	5. Mean of $P(t)$
2. Curl	4. Orientation	6. Variance of $P(t)$

- Features based on the normal vector field are:

- the average over the whole video sequence V of the divergence (scaling motion) and the curl (rotational motion) of the normal flow field,
- the peakiness of the normal flow field distribution, defined as the average flow magnitude divided by its standard deviation,
- the orientation homogeneity of the normal flow field: $\phi = \frac{\left\|\sum_{i\in\Omega} \boldsymbol{v}_N^i\right\|}{\sum_{i\in\Omega} \|\boldsymbol{v}_N^i\|} \in [0,1]$ where $\boldsymbol{v}_N^i$ is the normal flow vector at i and Ω is the set of points with nonzero normal flow vectors. ϕ reflects the flow homogeneity of the DT compared to its mean orientation. A detailed description of its meaning is given in [9].

- The two spatiotemporal regularity features are the temporal mean and variance of the maximum periodicity value $P(t)$ computed for the original greyscale sequence. Their computation on the normal flow sequence is under consideration.

All the selected features are *translation* and *rotation* invariant.

3 Results of Classification on a Real Dataset

3.1 The Dataset

The defined features have been applied on a dataset acquired by the MIT [11]. For test purposes, each sequence was divided into 8 non-overlapping subsets (samples), half in x, y and t. This dataset is browsing a wide range of possible DT occurrences (fig. 3): an escalator (A), a fire (B), a waving plastic sheet (C), clothes in a washing machine (D), a waving flag (E), smoke going upward (F), ripples on a river (G), a strong water vortex in toilet (H), trees (I) and boiling water (J).

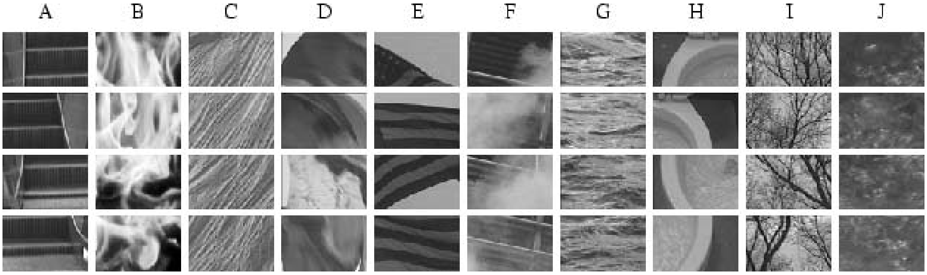
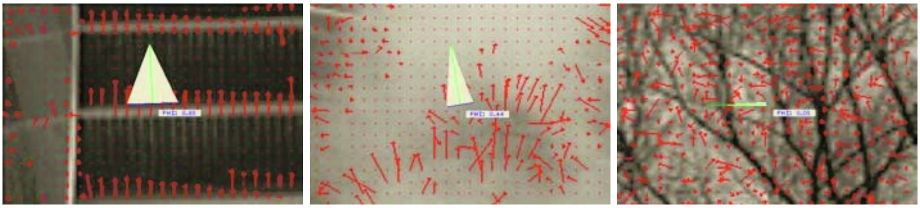


Fig. 3. The DT dataset (*courtesy of the MIT)*. Each column represents a class of dynamic texture and contains different samples of a class.

3.2 Experiment and Results

The 6 features have been computed for all the 8 samples of each of the 10 classes. Figure 4 illustrates the orientation homogeneity for class A, F and I.



(A) (F) (I)

Fig. 4. Orientation homogeneity for A ($\phi = 0.85$), F ($\phi = 0.44$) and I ($\phi = 0.05$). The main orientation is pointed by the triangle and its homogeneity is proportional to the base of this triangle.

For a rigid and well-oriented motion like the escalator A, the homogeneity value on orientation is high, reflecting a consistent main motion flow. The smoke F is not well segmented and has a lower value for orientation homogeneity. However, the ascending motion of the smoke is still extractable. The tree waving in the wind (class I) has a very low main orientation value due to its oscillating motion, resulting in an overall null displacement on the sequence.

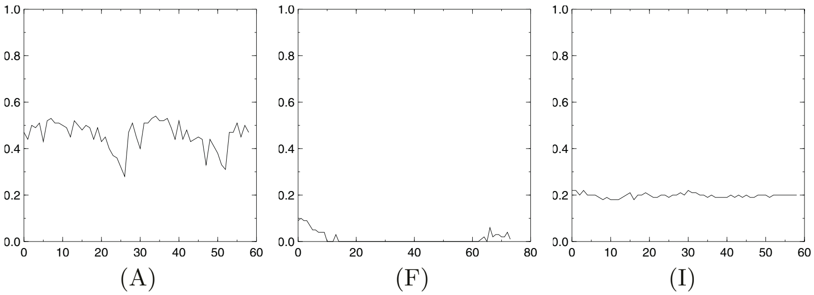


Fig. 5. Maximum regularity values for each frame. Mean values on the sequences are $\overline{P_A} = 0.460$, $\overline{P_F} = 0.013$ and $\overline{P_I} = 0.198$.

Figure 5 exemplifies the temporal evolution of the maximum periodicity value for the same DT. The size of the sliding window was set to 40 × 40 pixels. Dynamic textures A and I have significant and stable regularity values through time, whereas F appears as a random texture.

After normalization of the features, the leave-one-out classification test was carried out based on the nearest class. The test runs as follows. Given a class represented by 8 samples, one of the samples is selected. The other 7 samples are used as the learning samples and the mean values of the 6 features are computed for these 7 samples. The mean feature values for each of the other 9 classes are computed for all 8 samples. Then the distance between the selected sample s with feature vector $F_i^{(s)}$ and a class c represented by its mean feature vector $\overline{F_i^{(c)}}$ is computed as

$$D(s,c) = \sum_{i=1}^{6} w_i \cdot \|F_i^{(s)} - \overline{F_i^{(c)}}\|$$

where the weights w_i were set empirically as $w_i = 1$ except for $w_4 = 3$.

The sample s is classified as belonging to the nearest class n: $D(s,n) < D(s,c)$ for all $c \neq n$. This procedure is repeated for each sample of each class. Table 2 is the confusion matrix C_{pq} of the test.

C_{pq} shows the number of times a sample from class p was classified as belonging to class q, with the off-diagonal elements being misclassifications (see figure 3 for the order of the classes).

The overall accuracy is 93.8%. There are 5 cases of misclassification, with for instance 2 occurences of D (cloth in a laundry) were classified as E (waving flag).

This success rate has been obtained with only 6 features. Moreover, while the feature based on orientation homogeneity seems to be the most discriminative, each feature plays a role. For instance, the classification was performed with setting the regularity weights to zero ($w_5 = w_6 = 0$), and the success rate dropped to 85%. The regularity enables in this case to reduce the ambiguity between class D (laundry) and class E (flag): without the regularity features, there are 3 more cases of misclassification between those 2 classes.

Table 2. Confusion matrix C_{pq}.

True	A	B	C	D	E	F	G	H	I	J
A	**8**	0	0	0	0	0	0	0	0	0
B	0	**7**	0	1	0	0	0	0	0	0
C	0	0	**8**	0	0	0	0	0	0	0
D	0	1	0	**7**	0	0	0	0	0	0
E	0	0	0	2	**6**	0	0	0	0	0
F	0	0	1	0	0	**7**	0	0	0	0
G	0	0	0	0	0	0	**8**	0	0	0
H	0	0	0	0	0	0	0	**8**	0	0
I	0	0	0	0	0	0	0	0	**8**	0
J	0	0	0	0	0	0	0	0	0	**8**

4 Conclusion and Future Prospects

This article deals with the novel issue of dynamic texture recognition.

We have proposed a method for extracting quantitative features of dynamic textures. Based on the normal flow and on the texture regularity, the derived features are fast to compute and easily interpretable by a human operator. Their use in a full classification test has enabled to show their discriminative power: with only 6 features, the success rate reaches 93.8%.

Tests on a larger database will however be performed to better assess the capabilities and the limits of the proposed approach. Although the MIT database is currently the most frequently used collection of dynamic textures, no standard, rich dataset for comparing different DT classification algorithms seems to exist. It is obvious that the MIT dataset is not sufficient for statistically significant performance evaluation. However, some comparisons can still be done and certain conclusions can be drawn.

In particular, Szummer et al. [11] classify the MIT data based on the 8 best matches and obtain the accuracy of 95%. However, their approach needs much more computation than the proposed one, so their method does not seem to be applicable to DT based video retrieval, which is our ultimate goal.

Otsuka et al. [7] use only 4 of the 10 MIT sequences and obtain with 5 features a classification accuracy of 97.8%. Our method yields a similar result with a similar number of features, but for a significantly larger number of classes (all 10). Finally, Peh and Cheong [8] report on a classification accuracy of 87% achieved with 6 features for 10 DT classes different from the MIT data.

Our current work also aims at studying the multi-scale properties dynamic textures in space and time. In a longer term, we will also investigate the retrieval of dynamic textures in unsegmented scenes.

Acknowledgment

This work was carried out during the tenure of an ERCIM fellowship, and was also supported by the EU Network of Excellence MUSCLE (FP6-507752).Thanks also go to David Tschumperlé for its CImg library (http://cimg.sourceforge.net).

References

1. P. Bouthemy and R. Fablet. Motion characterization from temporal cooccurrences of local motion-based measures for video indexing. In *Int. Conf on Pattern Recognition, ICPR'98*, volume 1, pages 905–908, Brisbane, Australia, August 1998.
2. D. Chetverikov. Pattern regularity as a visual key. *Image and Vision Computing*, 18:pp. 975–986, 2000.
3. R. Deriche. Recursively Implementing the Gaussian and Its Derivatives. In *Proc. Second International Conference On Image Processing*, pages 263–267, Singapore, September 7-11 1992.
4. A. Divakaran. An overview of MPEG-7 motion descriptors and their applications. In W. Sharbek, editor, *CAIP 2001*, Lecture Notes in Computer Science 2124, pages 29–40, Warsaw, Poland, September 2001. Springer Verlag.
5. B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.
6. Randal C. Nelson and Ramprasad Polana. Qualitative recognition of motion using temporal texture. *CVGIP: Image Understanding*, 56(1):pp. 78–89, 1992.
7. K. Otsuka, T. Horikoshi, S. Suzuki, and M. Fujii. Feature extraction of temporal texture based on spatiotemporal motion trajectory. In *Int. Conf. on Pattern Recog. ICPR'98*, volume 2, pages 1047–1051, Brisbane, Australia, August 1998.
8. C.H. Peh and L.-F. Cheong. Synergizing spatial and temporal texture. *IEEE Transactions on Image Processing*, 11(10):pp. 1179–1191, 2002.
9. R. Péteri and D. Chetverikov. Qualitative characterization of dynamic textures for video retrieval. In *Proceedings ICCVG*, Warsaw, Poland, september 2004. To appear in Kluwer series on Computational Imaging and Vision.
10. P. Saisan, G. Doretto, Ying Nian Wu, and S. Soatto. Dynamic texture recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, volume 2, pages 58–63, Kauai, Hawaii, December 2001.
11. Martin Szummer and Rosalind W. Picard. Temporal texture modeling. In *Proc. IEEE International Conference on Image Processing*, volume 3, pages 823–826, 1996.
12. P. Wu, Y.M. Ro, C.S. Won, and Y. Choi. Texture descriptors in MPEG-7. In W. Sharbek, editor, *CAIP 2001*, Lecture Notes in Computer Science 2124, pages 21–28, Warsaw, Poland, September 2001. Springer Verlag.
13. J. Zhong and S. Scarlaroff. Temporal texture recongnition model using 3d features. Technical report, MIT Media Lab Perceptual Computing, 2002. 7 pages.

Detector of Image Orientation Based on Borda-Count

Loris Nanni and Alessandra Lumini

DEIS, IEIIT – CNR, Università di Bologna
Viale Risorgimento 2, 40136 Bologna, Italy
lnanni@deis.unibo.it

Abstract. Accurately and automatically detecting image orientation is a task of great importance in intelligent image processing. In this paper, we present automatic image orientation detection algorithms based on these features: color moments; harris corner; phase symmetry; edge direction histogram. The statistical learning support vector machines, AdaBoost, Subspace classifier are used in our approach as classifiers. We use Borda Count as combination rule for these classifiers. Large amounts of experiments have been conducted, on a database of more than 6,000 images of real photos, to validate our approaches. Discussions and future directions for this work are also addressed at the end of the paper.

1 Introduction

With advances in the multimedia technologies and the advent of the Internet, more and more users are very likely to create digital photo albums. Moreover, the progress in digital imaging and storage technologies have made processing and management of digital photos, either captured from photo scanners or digital cameras, essential functions of personal computers and intelligent home appliances. To input a photo into a digital album, the digitized or scanned image is required to be displayed in its correct orientation. However, automatic detection of image orientation is a very difficult task. Humans identify the correct orientation of an image through the contextual information or object recognition, which is difficult to achieve with present computer vision technologies.

However there are some external information that can be considered to improve the performance of a detector of image orientation in such a case where the acquisition source is known: a photo acquired by a digital camera often is taken in the normal way (i.e. 0° rotation), sometimes rotated by 90° or 270°, but very seldom rotated by 180°; a set of images acquired by digitalization of one film very unlikely has an orientation differing more than 90° (i.e. horizontal images are all straight or all upset), therefore the orientation of the most of the images belonging to the same class can be used to correct classification errors.

Since image orientation detection is a relatively new topic, the literature about this is quite sparse. In [1] a simple and rapid algorithm for medical chest image orientation detection has been developed. The most related work that investigates this problem was recently presented in [2][3][4][5]. In [5] the authors have extracted edge-based structural features, and color moment features: these two sources of information are incorporated into a recognition system to provide complimentary information for ro-

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 231–238, 2005.

bust image orientation detection. Support vector machines (SVMs) based classifiers are utilized. In [18] the authors propose an automated method based on the boosting algorithm to estimate image orientation.

The combination of multiple classifiers was shown to be suitable for improving the recognition performance in many difficult classification problems [10]. Recently a number of classifier combination methods, called ensemble methods, have been proposed in the field of machine learning. Given a single classifier, called the base classifier, a set of classifiers can be automatically generated by changing the training set [11], the input features [12], the input data by injecting randomness, or the parameters and architecture of the classifier. A summary of such methods is given in [13].

In this work, we assume that the input image is restricted to only four possible rotations that are multiples of 90° (figure 1). Therefore, we represent the orientation detection problem as a four-class classification problem (0°, 90°, 180°, 270°). We show that a combination of different classifiers trained in different feature spaces obtains a higher performance than a stand-alone classifier.

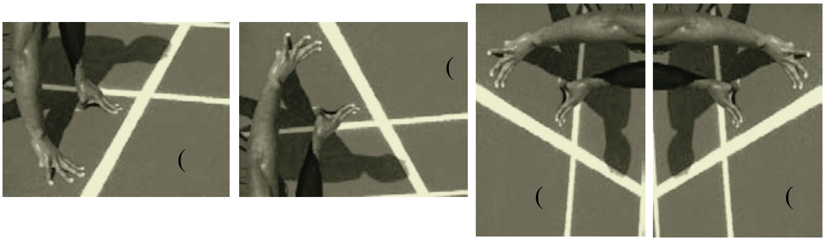


Fig. 1. The four possible orientations of acquisition for a digital image: (a) 0°, (b) 180°, (c) 90°, (d) 270°.

2 System Overview

In this section a brief description of the feature extraction methodologies, feature transformations, classifiers and ensemble methods combined and tested in this work is given.

2.1 Feature Extraction

Feature extraction is a process that extracts a set of new features from the original image representation through some functional mapping. In this task it is important to extract “local” features sensitive to rotation, in order to distinguish among the four orientations: for example the global histogram of an image is not a good feature because it is invariant to rotations. To overcome this problem an image is first divided in blocks, and then the selected features are extracted from each block. Several block decomposition have been proposed in the literature depending on the set of features to be extracted, in this work we adopt a regular subdivision in $N \times N$ non overlapping blocks (we empirically select N=10) and the features are extracted from these local regions.

2.1.1 Color Moments (COL)

It is shown in [14] that color moments of an image in the LUV color space are very simple yet very effective for color-based image analysis. We use the first order (mean color) and the second order moments (color variance) as our COL features to capture image chrominance information, so that for each block 6 COL features (3 mean and 3 variance values of L, U, V components) are extracted. Finally within each block, the COL vector is normalized such that the sum of each component square is one.

2.1.2 Edge Direction Histogram (EDH)

The edge-based structural features are employed in this work to capture the luminance information carried by the edge map of an image. Specifically, we utilize the edge direction histogram (EDH) to characterize structural and texture information of each block, similar as that in [14]. The Canny edge detector [15] is used to extract the edges in an image. In our experiments, we use a total of 37 bins to represent the edge direction histogram. The first 36 bins represent the count of edge points of each block with edge directions quantized at 10 intervals, and the last bin represents the count of the number of pixels that do not contribute to an edge (which is the difference between the dimension of a block and the first 36 bins).

2.1.3 Harris Corner Histogram (HCH)

A corner is a point that can be extracted consistently over different views, and there is enough information in the neighborhood of that point so that corresponding points can be automatically matched. The corner features are employed in this work to capture information about the presence of details in blocks by counting the number of corner points in each block. For corner detection we use the Harris corner detector [16].

2.1.4 Phase Symmetry (PHS)

Phase symmetry is an illumination and contrast invariant measure of symmetry in an image. These invariant quantities are developed from representations of the image in the *frequency domain*. In particular, phase data is used as the fundamental building block for constructing these measures. Phase congruency [17] can be used as an illumination and contrast invariant measure of feature significance. This allows edges, lines and other features to be detected reliably, and fixed thresholds can be applied over wide classes of images. Points of local symmetry and asymmetry in images can be detected from the special arrangements of phase that arise at these points, and the level of symmetry/asymmetry can be characterized by invariant measures. In this work we calculate the phase symmetry image [17] to count the number of symmetry pixels in each block.

The above COL, EDH, HCH and PHS vectors are normalized within each block. In order to accommodate the scale differences over different images during the feature extraction, all the features extracted are also normalized over training examples to the same scale. A linear normalization procedure has been performed, so that features are in the range [0,1].

2.2 Feature Transformation

Feature transformation is a process through which a new set of features is created from existing one. We adopt a Karhunen-Lòeve transformation (KL) to reduce the feature set to a lower dimensionality, maintaining only the most discriminant features. This step is performed since the original dimension of the feature space is too large to make training of classifiers feasible.

Given a F-dimensional data points $\mathbf{x}$, the goal of KL [7] is to reduce the dimensionality of the observed vector. This is obtained by finding k principal axes, denoted as principal component, which are given by the eigenvectors associated with the k largest eigenvalues of the covariance matrix of the training set.

In this paper the features extracted are reduced by KL to 100-dimensional vector.

2.3 Classifiers

A classifier is a component that uses the feature vector provided by the feature extraction or transformation to assign a pattern to a class. In this work, we test the following classifiers:

- AdaBoost (AB) [9];
- Polynomial-Support Vector Machine (P-SVM) [8];
- Radial basis function-Support Vector Machine (R-SVM) [8]
- Subspace (SUB) [6];

2.4 Multiclassifier Systems (MCS)

Multiclassifier systems are special cases where different approaches are combined to resolve the same problem. They combine output of various classifiers trained using different datasets by a Decision Rule [10]. Several decision rules can be used to determine the final class from an ensemble of classifiers; the most used are: Vote rule (vote), Max rule (max), Min rule (min), Mean rule (mean), Borda count (BORDA). The best classification accuracy for this orientation detection problem was achieved, in our experiments, using Borda Count.

For the Borda Count method, each class gets 1 point for each last place vote received, 2 points for each next-to-last point vote, etc., all the way up to N points for each first place vote (where N is the number of candidates/alternatives). The candidate with the largest point total wins the election.

2.5 Correction Rule

We implement a very simple heuristic rule that takes into account the acquisition information of an image to correct the classification response. After evaluating all the photos belonging to the same roll, we count the number of photos labelled as 0° and 180° and we select as correct orientation the one having the larger number of images, thus changing the labelling of images assigned to the wrong class. The same operation can be performed for the classes 90° an 270°.

3 Experimental Comparison

We carried out some tests to evaluate both the features extracted and the classifiers used. In order to test our correction rule we use a dataset of images acquired by analogical cameras and digitalized by roll film scanning. The dataset is composed by about 6,000 images from 350 rolls. This dataset is substantially more difficult than others tested in the literature, since it is composed by 80% indoor photos which are hardly classifiable: it is especially difficult to detect the orientations of indoor images (i.e. it is very hard to detect the orientations of a face) because we lack the discriminative features for indoor images, while for outdoor images there are lots of useful information which can be mapped to low-level features, such as sky, grass, building and water.

In figure 2 some images taken from our dataset are shown, where also outdoor images seem to be difficult to classify.

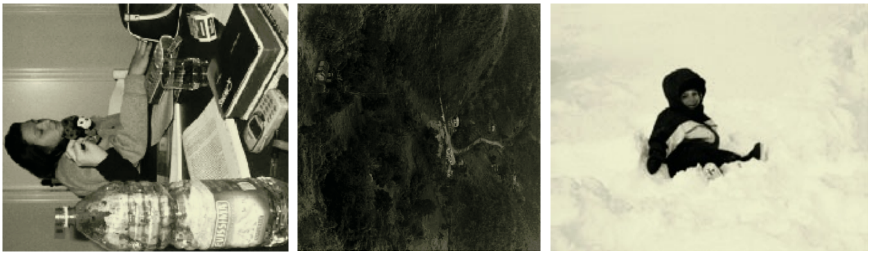


Fig. 2. Some images from our dataset which appear to be difficult to classify.

The classification results proposed in the following graphics are averaged on 10 tests, each time randomly resampling the test set, but maintaining the distribution of the images from different classes of orientation. The training set was composed by images taken from rolls not used in the test set, thus the test set is poorly correlated to the training data. For each image of the training set, we employ four features corresponding to four orientations (only one has to be extracted, the other three can be simply calculated).

We perform experimentations to verify that all the sets of features proposed are useful to improve the performance: the results in figure 3 show that the accuracy increases from 0.499 using only COL features to 0.557 using a combination of all the

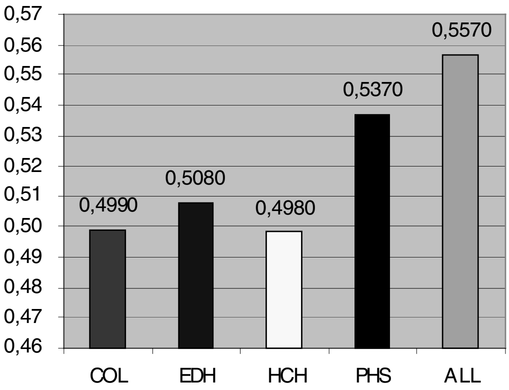


Fig. 3. Error rate using different feature spaces and R-SVM as classifier.

features. Moreover we show that the best performance using a single set of features are obtained using PHS, however the sequential combination of all the features (ALL) is better than PHS.

We perform experimentations to verify which classifiers is the best of the pool: the results in figure 4 show that R-SVM is the best "stand-alone" classifier, but BORDA (combining all the feature spaces and all the classifiers) is better than SVM.

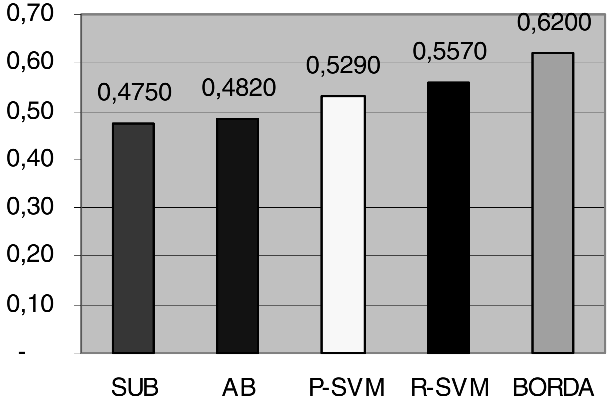


Fig. 4. Error rate using different classifiers.

The third experiment we carried out was aimed to compare the performance of the proposed method and the methods in [4][18]. We make a comparison at different level of rejection, adopting a very simple rejection rule: we reject the images for which the classifier has confidence lower than a threshold. In our multiclassifier BORDA, we use the confidence obtained by the best classifier (R-SVM with ALL as features). The performance denoted by BORDA-ROLL have been obtained by coupling to our method the correction rule described in section 2.5.

From the result shown in figure 5, we can see that with the same training data, the proposed fusion method performs better than the polynomial SVM trained using COL or COL+EDH features as proposed in [4][18]. In these papers the authors propose the following solutions to increase the performance of the standard SVM classifier:

- An AdaBoost method ([4]) where the feature space is obtained extending the original feature set (COL+EDH) by combining any two features with addition operation and thus getting a very large feature set of dimensions;
- A two layer SVMs ([4]) (with trainable combiner);
- A rejection scheme to reject more indoor than outdoor images at the same level of confidence score.

However these solutions grant only a slighter improvement of performance with respect to a standard SVM classifier, while adopting our fusion method the performance are really better than SVM. Moreover our correction rule has proven to be well suited for this problem, since it allows to improve the performance of the base method.

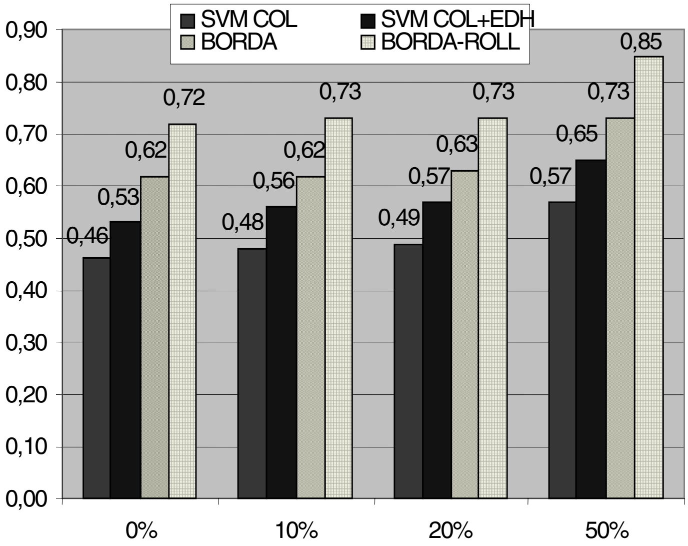


Fig. 5. Error rate using several classifiers at different levels of rejection.

4 Conclusions

We have proposed an automatic approach for content-based image orientation detection. Extensive experiments on a database of more than 6,000 images were conducted to evaluate the system. The experimental results, obtained on a dataset of "difficult" images which is very similar to real applications, show that our approach outperforms SVM and others "stand-alone" classifiers. However, the general image orientation detection is still a challenging problem. It is especially difficult to detect the orientations of indoor images because we lack the discriminative features for indoor images. The directions of our future work will concentrate on indoor images.

References

1. Evano, M.G., McNeill, K.M.: Computer recognition of chest image orientation, in: Proc. Eleventh IEEE Symp. on Computer-Based Medical Systems, (1998) 275–279.
2. Poz, A.P.D., Tommaselli, A.M.G.: Automatic absolute orientation of scanned aerial photographs, in: Proc. Internat. Symposium on Computer Graphics, Image Processing, and Vision, (1998) 295–302.
3. Vailaya, A., Zhang, H., Jain, A.K.: Automatic image orientation detection, in: Proc. Sixth IEEE Internat. Conf. on Image Processing, (1999) 600–604.
4. Vailaya, A., Jain, A.K.: Rejection option for VQ-based Bayesian classification, in: Proc. Fifteenth Internat. Conf. on Pattern Recognition, (2000) 48–51.
5. Wang, Y.M., Zhang, H.: Detecting image orientation based on low-level visual content, Computer Vision and Image Understanding, (2004) 328–346.
6. Oja, E.: .Subspace Methods of Pattern Recognition. Letchworth, England: Research Studies Press Ltd. 1983.
7. Duda, R. O., Hart, P. E., Stork, D. G.: Pattern Classification, Wiley, 2nd edition 2000.

8. Cristianini, N., Shawe-Taylor, J.: An introduction to Support vector machines and other kernel-based learning methods, Cambridge University Press, Cambridge, UK, 2000.
9. Viola, P., Jones, M.: Fast and robust classification using asymmtric AdaBoost and a detector cascade, In NIPS 14, 2002.
10. Kittler, J., Roli, F. (Eds.),: 1st Int. Workshop on Multiple Classifier Systems. Springer, Cagliari, Italy, 2000.
11. Breiman, L.: Bagging predictors. Mach. Learning (2), (1996) 123–140.
12. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell., (1998) 832–844.
13. Dietterich, T.G.:. Ensemble methods in machine learning. in: (Kittler and Roli, 2000). (2000) 1–15.
14. Jain, A.K., Vailaya, A.: Image retrieval using color and shape, Pattern Recognition 29, (1996) 1233–1244.
15. Canny, J.F.: A computational approach to edge detection, IEEE Trans. Pattern Anal. Mach. Intell. 8 (6), (1986) 679–698.
16. Harris, C., Stephens, M.: A combined corner and edge detector, Fourth Alvey Vision Conference, (1988) 147-151.
17. Peter, K.: Image Features From Phase Congruency. Videre: A Journal of Computer Vision Research. MIT Press. Volume 1, Number 3, Summer 1999.
18. Zhang, L., Li, M., Zhang, H.: Boosting Image Orientation Detection with Indoor vs. Outdoor Classification, Sixth IEEE Workshop on Applications of Computer Vision, 2002.

Color Image Segmentation Using Acceptable Histogram Segmentation

Julie Delon[1], Agnes Desolneux[2], Jose Luis Lisani[3], and Ana Belen Petro[3]

[1] CMLA, ENS Cachan, France
julie.delon@cmla.ens-cachan.fr
[2] MAP5, Université Paris 5, France
desolneux@math-info.univ-paris5.fr
[3] Univ. Illes Balears, Spain
{joseluis.lisani,anabelen.petro}@uib.es

Abstract. In this paper, a new method for the segmentation of color images is presented. This method searches for an acceptable segmentation of 1D-histograms, according to a “monotone” hypothesis. The algorithm uses recurrence to localize all the modes in the histogram. The algorithm is applied on the hue, saturation and intensity histograms of the image. As a result, an optimal and accurately segmented image is obtained. In contrast to previous state of the art methods uses exclusively the image color histogram to perform segmentation and no spatial information at all.

1 Introduction

Image segmentation refers to partitioning an image into different regions that are homogenous with respect to some image feature. Thought most attention on this field has been focused on gray scale images, color is a powerful feature that can be used for image segmentation.

Among the classical techniques for color images segmentation, pixel-based techniques do not consider spatial information. The simplest pixel-based technique for segmentation is histogram thresholding which assumes that the histogram of an image can be separated into as many peaks (modes) as different regions are present in the image.

The existing techniques for histogram thresholding can be distinguished by the choice of the color component from which the histogram is obtained and by the modes extraction criterion. Concerning the first of these issues, some approaches ([7]) consider 3D histograms, that simultaneously contain all the color information in the image. However, storage and processing of multidimensional histograms is computationally expensive. For this reason most approaches consider 1D histograms computed for one or more color components in some color space (see for example [5] and [6]).

With respect to the modes extraction criterion, most methods are based on parametric approaches. That is, they assume the histogram to be composed of k random variables of a given distribution, for instance the Gaussian distribution, with different averages and variances. However, these methods require an estimation of the number of modes in the final segmentation and, moreover, the found modes have not proven to be relevant.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 239–246, 2005.

In this work, we present an automatic method for color image segmentation based on the detection of "menaingul modes" in histograms. We choose the HSI color space for the application of our approach to color images. This space has the advantage of separating color from intensity information.

The paper is structured as follows. The basic ideas of the method are exposed in the next section; in section 3, the application to color segmentation is presented; section 4 is devoted to display and comment some results on color image segmentation; the conclusions are presented in the final section.

2 Histogram Analysis by Helmholtz Principle

In 2003, A. Desolneux, L. Moisan and J.M. Morel ([3]) defined a new parameter-free method for the detection of meaningful events in data. An event is called ε-meaningful if its expectation under the a contrario random uniform assumption is less than ε. Let us state what this definiton yields in the case of the histogram modes.

2.1 Uniform Hypothesis. Meaningful Intervals and Meaningful Gaps of a Histogram

We will consider a discrete histogram r, that is N points distributed on L values, $\{1, ..., L\}$. For each discrete interval $[a, b]$ of $\{1, ..., L\}$, $r(a, b)$ will represent the proportion of points in the interval. For each interval $[a, b]$ of $\{1, ..., L\}$, we note $p(a, b) = \frac{b-a+1}{L}$ the relative length of the interval. The value $p(a, b)$ is also, under the uniform assumption, the probability for a point to be in $[a, b]$. Thus, the probability that $[a, b]$ contains at least a proportion $r(a, b)$ of points among the N is given by the binomial tail $\mathcal{B}(N, Nr(a, b), p(a, b))$, where $\mathcal{B}(n, k, p) = \sum_{j=k}^{n} \binom{n}{j} p^j (1-p)^{n-j}$. The number of false alarms of $[a, b]$ is:

$$NFA([a, b]) = \frac{L(L+1)}{2} \mathcal{B}(N, Nr(a, b), p(a, b)).$$

Thus, an interval $[a, b]$ is said ε-meaningful if it contains "more points" than the expected average, in the sense that $NFA([a, b]) \leq \varepsilon$, that is

$$\mathcal{B}(N, Nr(a, b), p(a, b)) < \frac{2\varepsilon}{L(L+1)}.$$

In the same way, an interval $[a, b]$ is said to be an ε-meaningful gap if it contains "less points" than the expected average.

If ε is not too large (in practice, we will always use $\varepsilon = 1$) an interval cannot be at the same time an ε-meaningful interval and an ε-meaningful gap. Now, these binomial expressions are not always easy to compute, especially when N is large. In practice, we adopt the large deviation estimate to define meaningful intervals and gaps.

Definition 1. *The relative entropy of an interval* $[a, b]$ *(with respect to the prior uniform distribution p) is defined by*

$$H([a, b]) = r(a, b) \log \frac{r(a, b)}{p(a, b)} + (1 - r(a, b)) \log \frac{1 - r(a, b)}{1 - p(a, b)}.$$

$H([a,b])$ is the Kullback-Kleiber distance between two Bernoulli distributions of respective parameters $r(a,b)$ and $p(a,b)$ ([2]), that is $H([a,b]) = \mathrm{KL}(r(a,b)||p(a,b))$.

Definition 2. *An interval $[a,b]$ is said to be an **ε-meaningful interval** (resp. **ε-meaningful gap**) if $r(a,b) \geq p(a,b)$ (resp. $r(a,b) \leq p(a,b)$) and if*

$$H([a,b]) > \frac{1}{N} \log \frac{L(L+1)}{2\varepsilon}$$

2.2 Monotone Hypothesis

How can we detect meaningful intervals or gaps if we know that the observed objects follow a non-uniform distribution (e.g. decreasing or increasing)? We want now to define the meaningfulness of an interval with respect to the decreasing hypothesis (the definitions and results for the increasing hypothesis can be deduced by symmetry). We will call $\mathcal{D}(L)$ the space of all decreasing densities on $\{1,2,...,L\}$ and $\mathcal{P}(L)$ be the space of normalized probability distributions on $\{1,2,...,L\}$, *i.e.* the vectors $r = (r_1,...,r_L)$ such that: $\forall i \in \{1,2,...,L\}, r_i \geq 0$ and $\sum_{i=1}^{L} r_i = 1$.

If $r \in \mathcal{P}(L)$ is the normalized histogram of our observations, we need to estimate the density $p \in \mathcal{D}(L)$ in regards to which the empirical distribution r has the "less meaningful" gaps and intervals, which is summed up by the optimization problem

$$\tilde{r} = argmin_{p \in \mathcal{D}(L)} \min_{[a,b] \in \{1,2,...,L\}} KL(r(a,b)||p(a,b)).$$

The meaningfulness of intervals and gaps can then be defined relatively to this distribution $\tilde{r}$. Note that the uniform distribution is a particular case of decreasing density, which means that this formulation strengthens the previous theory: if there is no meaningful interval or gap in regards to the uniform hypothesis, there will be no meaningful interval or gap in regards to the decreasing hypothesis.

However, this optimization problem is uneasy to solve. We choose to slightly simplify it by approximating $\tilde{r}$ by the Grenander estimator $\bar{r}$ of r ([4]), which is defined as the nonparametric maximum likelihood estimator restricted to decreasing densities on the line.

Definition 3. *The histogram $\bar{r}$ is the unique histogram which achieves the minimal Kullback-Leibler distance from r to $\mathcal{D}(L)$, i.e. $KL(r||\overline{r}) = \min_{p \in \mathcal{D}(L)} KL(r||p)$.*

It has been proven ([1]) that $\overline{r}$ can easily be derived from r by an algorithm called "Pool Adjacent Violators" that leads to a unique decreasing step function $\bar{r}$.

Pool Adjacent Violators
Let $r = (r_1,...,r_L) \in \mathcal{P}$ be a normalized histogram. We consider the operator $D : \mathcal{P} \to \mathcal{P}$ defined by: for $r \in \mathcal{P}$, and for each interval $[i,j]$ on which r is increasing, i.e. $r_i \leq r_{i+1} \leq ... \leq r_j$ and $r_{i-1} > r_i$ and $r_{j+1} < r_j$, we set

$$D(r)_k = \frac{r_i + ... + r_j}{j-i+1} \text{ for } k \in [i,j], \text{ and } D(r)_k = r_k \text{ otherwise.}$$

This operator D replaces each increasing part of r by a constant value (equal to the mean value on the interval).

After a finite number (less than the size L of r) of iterations of D we obtain a decreasing distribution denoted $\overline{r}$, $\overline{r} = D^L(r)$.

An example of histogram and its Grenander estimator is shown on Figure 1.

Now, the definitions of meaningful interval gaps are analogous to the ones introduced in the uniform case, the uniform prior being just replaced by the global decreasing estimate $\overline{r}$ of the observed normalized histogram r.

Definition 4. *Let r be a normalized histogram. We say that an interval $[a, b]$ is ε-meaningful for the decreasing hypothesis (resp. an ε-meaningful gap for the decreasing hypothesis) if $r(a, b) \geq \overline{r}(a, b)$ (resp. $r(a, b) \leq \overline{r}(a, b)$) and*

$$H_{\overline{r}}([a, b]) > \frac{1}{N} \log \frac{L(L+1)}{2\varepsilon},$$

where $H_{\overline{r}}([a, b]) = \mathrm{KL}(r(a, b)||\overline{r}(a, b))$.

We are now able to define precisely what we called "to be almost decreasing on a segment".

Definition 5. *We say that a histogram **follows the decreasing (resp. increasing) hypothesis** on an interval if it contains no meaningful gap for the decreasing (resp. increasing) hypothesis on the interval.*

2.3 Acceptable Segmentations

The aim of our segmentation is to split the histogram in separated "modes". We will call "mode" an interval on which the histogram follows the increasing hypothesis on a first part and the decreasing one on the second part.

Definition 6. *We say that a histogram r **follows the unimodal hypothesis** on the interval $[a, b]$ if it exists $c \in [a, b]$ such that r follows the increasing hypothesis on $[a, c]$ and r follows the decreasing hypothesis on $[c, b]$.*

Such a segmentation exists. Indeed, the segmentation defined by all the minima of the histogram as separators follows obviously the unimodal hypothesis on each segment. But if there are small fluctuations it is clear that it is not a reasonable segmentation (see Fig. 2 left). We present a procedure that finds a segmentation much more reasonable than the segmentation defined by all the minima. We want to build a minimal (in terms of numbers of separators) segmentation, which leads us to introduce the notion of "acceptable segmentation".

Definition 7. *Let r be a histogram on $\{1, ..., L\}$. We will say that a segmentation s of r is **acceptable** if it verifies the following properties:*

- *r follows the unimodal hypothesis on each interval $[s_i, s_{i+1}]$.*
- *there is no interval $[s_i, s_j]$ with $j > i + 1$, on which r follows the unimodal hypothesis.*

The two requirements allow us to avoid under-segmentations and over-segmentations, respectively. It is clear in the discrete case that such a segmentation exists: we can start with the limit segmentation containing all the minima of r and gather the consecutive intervals together until both properties are verified. It is the principle used in the next algorithm:

Fine to Coarse (FTC) Segmentation Algorithm:

1. *Define the finest segmentation (i.e. the list of all the minima)* $S = \{s_1, \ldots, s_n\}$ *of the histogram.*
2. *Repeat:*
   Choose i *randomly in* $[2, length(S) - 1]$. *If the modes on both sides of* s_i *can be gathered in a single interval* $[s_{i-1}, s_{i+1}]$ *following the unimodal hypothesis, group them. Update* S.
   Stop when no more unions of successive intervals follows the unimodal hypothesis.
3. *Repeat step 2 with the unions of* j *intervals,* j *going from* 3 *to* $length(S)$.

A result of this algorithm is shown on Figure 2. The left part of the figure shows the initialization of the algorithm (all the minima), and the final result is on the right.

Now that we have set up an algorithm which ensures the construction of an acceptable segmentation, we will devote the next section to the application of the proposed method to color image segmentation.

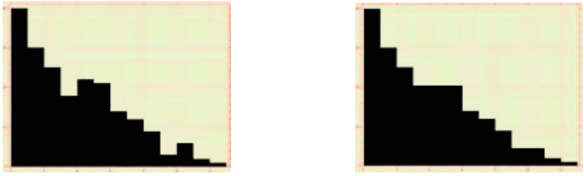


Fig. 1. The right histogram is the Grenander estimator of the left histogram, computed by the "Pool Adjacent Violators" algorithm. Observe that the new histogram is a decreasing function.

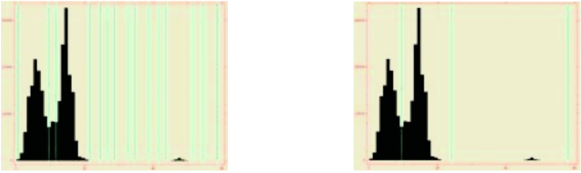


Fig. 2. Left: all the minima of the histogram. Right: remaining minima after the fine to coarse algorithm.

3 Color Image Segmentation

We apply the FTC algorithm on the hue histogram of a color image, in order to obtain a first segmentation. At this step a lot of color information, contained in the saturation and intensity components, has been lost. Then, we follow the same process by applying the algorithm to the saturation and intensity histograms of each mode obtained previously.

For the practical implementation of the algorithm, we must take into account that, in the discrete case, a quantization problem appears when we try to assign hue values to quantized color points in the neighborhood of the grey axis. A solution to this problem is to discard points that have saturation smaller than $\frac{Q}{2\pi}$, where Q is the number of quantized hue values. This requirement defines a cylinder in the HSI color space called the *grey cylinder*, since all the points contained in it will be considered as grey values.

Our algorithm can be described by the following steps:

Color Image Segmentation Algorithm

1. *Apply the FTC algorithm on the hue histogram of the image. Let S be the obtained segmentation.*
2. *Link each pixel of the grey cylinder to its corresponding interval $S_i = [s_i, s_{i+1}]$, according to its hue value.*
3. *For each i, construct the saturation histogram of all the pixels of the image whose hue belongs to S_i. Do not take into account the pixels of the grey cylinder. Apply the FTC algorithm on the corresponding saturation histogram. For each i, let $\{S_{i,i_1}, S_{i,i_2}, \ldots\}$ be the obtained segmentation.*
4. *For each i, link each pixel of the grey cylinder which belonged to the interval S_i, to the lower saturation interval $S_{i,i1}$ obtained in the previous segmentation step.*
5. *For each i and each j, compute and segment the intensity histogram of all the pixels whose hue and saturation belong to S_i and S_{i,i_j}, including those of the grey cylinder.*

It is also worth noting that the hue histogram is circular, which means that the hue value $0°$ is identified to the hue value $360°$.

4 Results

For each experiment we show five images: the first one is the original image; the second one corresponds to the hue histogram with the obtained modes marked with a dashed line between them; the third one is the segmented image after the application of the algorithm to the hue histogram; the fourth one corresponds to the segmented image after the application of the algorithm to the histograms of hue and saturation; and, finally, the segmented image after the application of the algorithm to the histograms of hue, saturation and intensity. The segmented images are displayed with the different modes represented by the mean values of the hue, saturation and intensity of all the pixels in the mode.

Figure 3 is the “ladybug” image in which we distinguish three different colors corresponding to different objects: background, leaf and ladybug. After applying the proposed separation approach to the hue histogram, we obtain three different modes, which correspond to the three referred objects. It is clear that the modes are detected independently of their relative number of pixels. Detection only depends on the meaningfulness of the mode, allowing the detection of small objects, as in the present example. The total number of colors in the final segmentation is 11, because of the great variety of shades in the background.

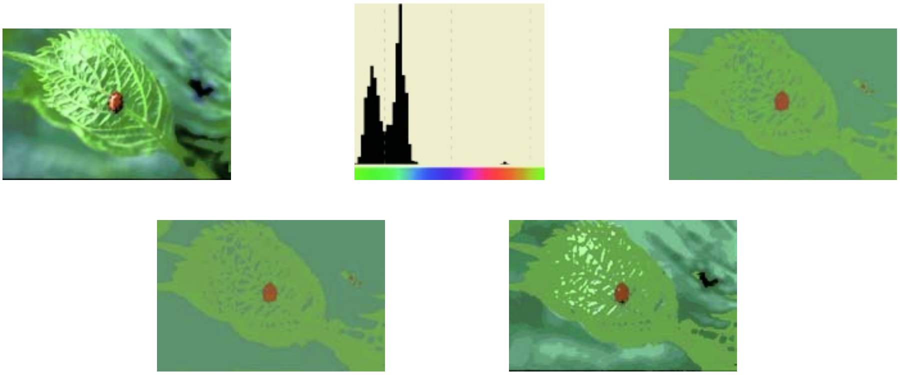


Fig. 3. The results of the proposed approach: Original image "ladybug"; hue histogram with the three obtained modes; resulting image with 3 colors after hue segmentation; resulting image with 4 colors after hue and saturation segmentation; resulting image with 11 colors after hue, saturation and intensity segmentation.

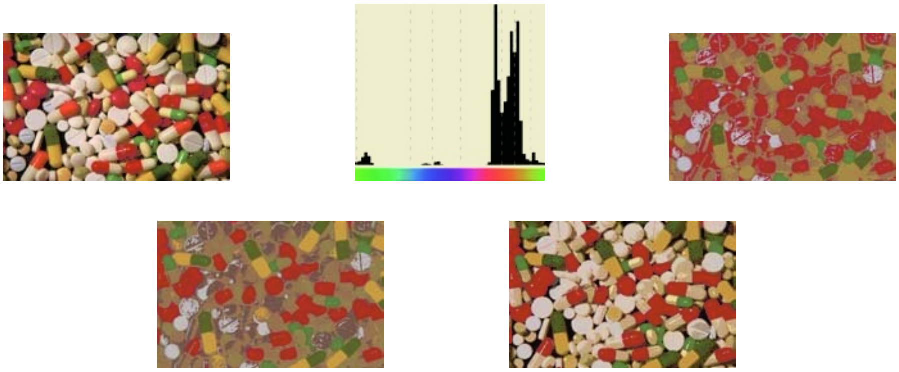


Fig. 4. The results of the proposed approach: Original image "pills"; hue histogram with the seven obtained modes; resulting image with 7 colors after hue segmentation; resulting image with 9 colors after hue and saturation segmentation; resulting image with 23 colors after hue, saturation and intensity segmentation.

In figure 4 we obtain seven modes in the first step of the segmentation. These colors correspond to two different greens, two different brows, one red, one blue and one cyan, which are clearly distinguished in the original image. Then, the saturation step add two new colors and finally we obtain twenty-three colors.

Figure 5 displays a third experiment. See figure caption for details.

5 Conclusion

In this paper, we have presented a new histogram thresholding method for color image segmentation. The method searches the optimal separators in the hue, saturation and intensity histograms of the color image. As a result, we obtain different modes which correspond to different regions in the image. This permits to segment the color image, by representing the different regions by their mean color.

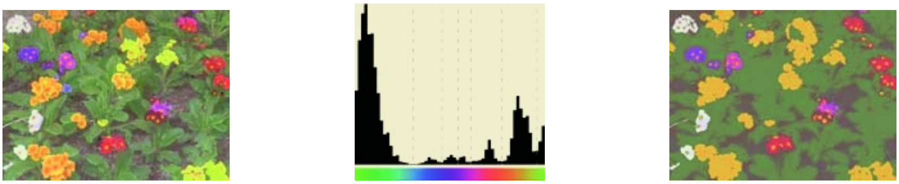


Fig. 5. The results of the proposed approach: Original image “flowers”; hue histogram with the six obtained modes; resulting image with 11 colors after hue, saturation and intensity segmentation.

As an application of the proposed method, we aim at a color analysis algorithm giving automatically, for every color image, a small, consistent and accurate list of the names of colors present in it, thus yielding an automatic color recognition system.

Acknowledgements

The authors gratefully thank F. Mettini for her kind authorization to use the “ladybug” image. J.L. Lisani and A. B. Petro thank the support by CICYT project TIC2002-02172.

References

1. M. Ayer, H.D. Brunk, G.M. Ewing, W.T. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26(4):641–647, 1955.
2. Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley Series in Telecommunications. John Wiley & Sons Inc., New York, 1991. A Wiley-Interscience Publication.
3. Agnès Desolneux, Lionel Moisan, and Jean-Michel Morel. A grouping principle and four applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4):508–513, 2003.
4. U. Grenander. On the theory of mortality measurement, part II. *Skand. Akt.*, 39:125–153, 1956.
5. K. Price R. Ohlander and D.R. Reddy. Picture segmentation using a recursive region splitting method. *Computer Graphics and Image Processing*, 8:313–333, 1978.
6. S. Tominaga. Color classification of natural color images. *Color Research and Application*, 17(4):230–239, 1992.
7. A. Trémeau, C. Fernandez-Maloigne, and P. Bonton. *Image numérique couleur, de l’acquisition au traitement*. Dunod, 2004.

Adding Subsurface Attenuation to the Beckmann-Kirchhoff Theory

Hossein Ragheb[1] and Edwin R. Hancock[2]

[1] Department of Computer Engineering, Bu-Ali Sina University,
Hamedan, Iran, PO Box 65175-4161
ragheb@basu.ac.ir

[2] Department of Computer Science, University of York,
York, YO1 5DD, UK
erh@cs.york.ac.uk

Abstract. In this paper we explore whether the Fresnel term can be used to improve the predictions of the Beckmann-Kirchhoff (B-K) model for moderately-rough surfaces. Our aim in developing this model is to describe subsurface scattering effects for surfaces of intermediate roughness. We use the BRDF measurements from the CUReT database to compare the predictions of the Fresnel correction process with several variants of the B-K model and the Oren and Nayar Model. The study reveals that our new Fresnel correction provides accurate predictions, which are considerably better than those achieved using both the alternative variants of the B-K model and the Oren-Nayar model.

1 Introduction

The modelling of rough surface reflectance is important in both computer vision and computer graphics, and has been the subject of sustained research activity for some four decades. The quest for a reflectance model that can accurately account for observed surface radiance under a variety of roughness conditions, and a variety of viewing geometries has proved to be an elusive one. Surface roughness can be characterized in a number of ways. For very-rough surfaces, one approach is to use a model which describes the distribution of surface wall cavity angles [10]. For rough surfaces which present a shiny appearance, roughness can be modelled using the angular distribution of microfacets [12]. An alternative that can capture both effects is to describe the roughness phenomenon using the variance and the correlation of variations in the surface height distribution [13].

Broadly speaking there are three approaches to the modelling of rough surface reflectance. The first of these is the purely empirical approach which has proved effective in the computer graphics domain for developing computationally efficient methods for synthesizing subjectively realistic surface appearance [11]. A more sophisticated approach is the so-called semi-empirical one which attempts to account for the main phenomenology of the light scattering process whilst falling short of developing a detailed physical model [10, 13, 15]. Finally, if a physically detailed model is required then wave scattering theory can be applied to understand the interaction of light with the surface relief distribution. Some of the earliest work aimed at modelling surface reflectance as a wave

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 247–254, 2005.

scattering process was undertaken by Beckmann [1]. Here the Kirchhoff integral is used to account for the scattering of light from rough surfaces characterized using the variance and correlation length of the surface relief distribution. The model is mathematically quite complex, and is hence not well suited for analysis tasks of the type encountered in computer vision. In particular, it is not a simple modification of Lambert's law. However, He et al. [5] exploited an improved representation of the Kirchhoff integral for surface synthesis in computer graphics. They proposed a comprehensive model that incorporates complex factors including surface statistics, sub-layer scattering, and polarization.

Unfortunately, the Beckmann model fails to account for the observed radiance at large scatter angles due to energy absorbtion, self shadowing and multiple scattering effects. Some of the problems have recently been overcome by Vernold and Harvey [14] who have used a simple Lambertian form factor to modify Beckmann's predictions. There is considerable debate about the meaning and proper use of the form factor [7, 8]. A number of alternatives have been used in the wave scattering literature. Based on boundary condition considerations Ogilvy [9] argues that it should be proportional to the cosine of the angle of light incidence. Nieto and Garcia [8] have suggested a different form that ensures energy normalization. Finally, it is important to note that the B-K model provides closed-form solutions only for slightly-rough and very-rough surfaces.

Our aim is to fill this gap in the literature by developing an extension of the Beckmann model that can be applied to surfaces of intermediate roughness. To do this we borrow ideas from the semi-empirical modelling of rough surface reflectance, where for slightly-rough or shiny surfaces, Wolff [15] has accounted for refractive attenuation in the surface-air layer by multiplying Lambert's cosine law by the Fresnel term. Our new variant of the Beckmann model incorporates subsurface refractive attenuation of light prior to wave scattering by multiplying the Kirchhoff scattering kernel by a form factor that depends on the Fresnel reflection coefficient. We compare the new model with a number of model variants suggested by B-K theory on BRDF measurements from the CUReT database [3]. The new model outperforms the alternatives studied and provides remarkably good fits to the CUReT data for surfaces of intermediate roughness.

2 Kirchhoff Scatter Theory

The Beckmann-Kirchhoff (B-K) theory attempts to account for the wave interactions of light with rough surfaces. We are interested in a surface illuminated by a parallel beam of light of known wavelength λ and viewed by a camera which is sufficiently distant from the surface so that perspective effects may be ignored. The incident light has zenith angle θ_i and azimuth angle ϕ_i with respect to the surface normal, while the zenith and azimuth angles of the viewer (scattering) direction with respect to the surface normal are θ_s and ϕ_s. The radiance of the incident light-beam at the location on the surface is $\mathcal{L}_i(\theta_i, \phi_i)$ and the outgoing radiance is $\mathcal{L}_o(\theta_i, \phi_i, \theta_s, \phi_s)$. If $\nu(\theta_i, \phi_i, \theta_s, \phi_s)$ is the bidirectional reflectance distribution function (BRDF) for the surface, we can write

$\mathcal{L}_o(\theta_i, \phi_i, \theta_s, \phi_s) \propto \nu(\theta_i, \phi_i, \theta_s, \phi_s)\mathcal{L}_i(\theta_i, \phi_i) \cos\theta_i d\omega$. We will make use of this relationship to compare the predictions of light scattering theory with BRDF data from the CUReT database [3]. In the CUReT database, the BRDF data have been tabulated by assuming that each surface radiance value is related to the corresponding image pixel brightness value by a gain and an offset. For each surface sample, we normalize the scattered radiance values by dividing them by the maximum radiance value. This ensures that the comparisons are accurate. The CUReT database provides BRDF measurements at 205 different illumination configurations for each surface sample.

The B-K model has two main physical parameters. The first of these is the root-mean-square (RMS) height deviation of the topographic surface features about the mean surface level which is denoted by σ. The height values are generally measured at equally spaced digitized data points. The height variations Δz of the surface are assumed to follow the Gaussian distribution function $W(\Delta z) = (1/\sigma\sqrt{2}) \exp(-\Delta z^2/2\sigma^2)$. The second parameter is the correlation length T which is defined in terms of the surface correlation function $C(\tau)$ where τ is the distance (length) parameter so that $T^2 = 2\int_0^\infty \tau C(\tau)d\tau$. The correlation function characterizes the random nature of a surface profile and the relative spacing of peaks and valleys. For a purely random surface, $C(\tau)$ decreases monotonously from its maximum value $C(0) = 1$ to $C(\infty) = 0$. Specifically, the correlation length is the lag-length at which the Gaussian correlation function $C(\tau) = \exp(-\tau^2/T^2)$ drops to $1/e$. Similarly, for the exponential correlation function $C(\tau) = \exp(-|\tau|/T)$ it follows that $C(T) = 1/e$.

According to the B-K model [1], for a surface that is slightly or moderately rough, the mean scattered power is

$$P(\theta_i, \phi_i, \theta_s, \phi_s) = \rho_0^2 e^{-g(\theta_i, \theta_s)} + \mathcal{D}(\theta_i, \phi_i, \theta_s, \phi_s) \tag{1}$$

where the first term is the coherent scattering component in the specular direction and the second term is the incoherent component due to diffuse scattering. In this paper we are only interested in reflectance modelling for surfaces that are moderately rough or very rough. For such surfaces the coherent scattering component in the specular direction is negligible, and $P(\theta_i, \phi_i, \theta_s, \phi_s) = \mathcal{D}(\theta_i, \phi_i, \theta_s, \phi_s)$. The B-K model provides an infinite series solution for both slightly-rough and moderately-rough surfaces [1]. When the correlation function is *Gaussian*, then with the geometry outlined above on the tangent plane (the incident beam has azimuth angle $\phi_i = \pi$) the diffuse component is

$$\mathcal{D}(\theta_i, \phi_i = \pi, \theta_s, \phi_s) = \frac{\pi T^2 F_{BK}^2(\theta_i, \theta_s, \phi_s)}{A \exp\left[g(\theta_i, \theta_s)\right]} \sum_{n=1}^{\infty} \frac{g^n(\theta_i, \theta_s)}{n!n} \exp\left[\frac{-T^2}{4n} v_{xy}^2(\theta_i, \theta_s, \phi_s)\right] \tag{2}$$

where $v_x(\theta_i, \theta_s, \phi_s) = k(\sin\theta_i - \sin\theta_s \cos\phi_s)$, $v_y(\theta_s, \phi_s) = -k(\sin\theta_s \sin\phi_s)$, $v_z(\theta_i, \theta_s) = -k(\cos\theta_i + \cos\theta_s)$, $v_{xy}^2(\theta_i, \theta_s, \phi_s) = v_x^2(\theta_i, \theta_s, \phi_s) + v_y^2(\theta_s, \phi_s)$ and $k = 2\pi/\lambda$. The scattering takes place from a rectangular surface patch of area $A = 4XY$ whose dimensions are $2X$ and $2Y$ in the X_0 and Y_0 directions. The quantity F_{BK} appearing in the diffuse component is the *geometric factor*, and its choice is of critical importance to the B-K model. The quantity

$g(\theta_i, \theta_s) = \sigma^2 v_z^2(\theta_i, \theta_s)$ plays an important role, since it has been used in the literature to divide surfaces into three broad categories. These are a) slightly-rough ($g(\theta_i, \theta_s) \ll 1$), b) moderately-rough ($g(\theta_i, \theta_s) \simeq 1$) and c) very-rough ($g(\theta_i, \theta_s) \gg 1$) surfaces.

The geometrical factor F which is derived by Beckmann (F_{BK}) is given by

$$F_{BK}(\theta_i, \theta_s, \phi_s) = \frac{1 + \cos\theta_i \cos\theta_s - \sin\theta_i \sin\theta_s \cos\phi_s}{\cos\theta_i(\cos\theta_i + \cos\theta_s)} \tag{3}$$

Unfortunately, as highlighted by several authors [14], this choice fails to reliably predict the scattering behavior at large angles of incidence and scattering. To overcome this problem based on phenomenological arguments Vernold and Harvey [14] argue for the use the geometrical factor $F_{VH}^2 = \cos\theta_i$ that is Lambertian in form and depends only on the cosine of the incidence angle. This modification gives reasonable experimental agreement with scattering data for rough surfaces at large angles of incidence and large scattering angles.

There is considerable debate about the meaning and proper use of the geometrical (inclination) factor F [8]. In fact, a variety of forms for F have been proposed in the wave scattering literature. For instance, based on boundary condition considerations Ogilvy [9] argues for the factor $F_Y = F_{BK} \cos\theta_i$. Ogilvy also stresses that the F-factor is independent of the choice of total or scattered field within the B-K integrand [9]. Note that the expression derived by Ogilvy is for the scattered intensity. To derive the scattered radiance, one should divide the scattered intensity by $\cos\theta_s$ [4]. Nieto and Garcia [8] have shown that the factor used by Kirchhoff is related to that of Beckmann by the formula $F_{NG} = F_{BK} \cos\theta_i / \cos\theta_s$, and note that a factor $\cos\theta_s / \cos\theta_i$ is necessary to ensure energy normalization. An atomic scattering model is used to derive their expressions, and hence it is not guaranteed that their model can predict macroscopic surface scattering effects accurately.

Here we follow Harvey et al. [4] and interpret Beckmann's result as diffracted scattered radiance rather than scattered intensity. This means that for moderately-rough and very-rough surfaces we can write $\mathcal{L}_o(\theta_i, \phi_i, \theta_s, \phi_s) \propto D(\theta_i, \phi_i, \theta_s, \phi_s)$.

3 Subsurface Refractive Attenuation

The Fresnel coefficient has been widely used to account for subsurface scattering. For instance, Wolff [15] has used it to correct Lambert's law for smooth surfaces. Torrance and Sparrow [12] have included the Fresnel term in their specular intensity model. It is also used in the more complex model of He et al. [5] which attempts to account for a number of effects including subsurface scattering.

Wolff has developed a physically motivated model for diffuse reflectance from smooth dielectric surfaces [15]. The model accounts for subsurface light scattering, using the Fresnel attenuation term. The attenuation term modifies Lambert's law in a multiplicative manner and accounts for the refractive attenuation in the surface-air layer. The model successfully accounts for a number of distinctive features of diffuse reflectance from smooth surfaces. For instance, it accounts for both diffuse reflectance maxima when the angles between the viewer and the light-source exceeds 50 degrees, and also significant departures from Lambert's law. According to this model, the surface radiance is given by

$$\mathcal{L}_o(\theta_i, \theta_s, n) = \varrho \mathcal{L}_i \cos\theta_i [1 - f(\theta_i, n)]\{1 - f(\sin^{-1}[(\sin\theta_s)/n], 1/n)\} \tag{4}$$

The attenuation factor, $0 \leq f(\alpha_i, n) \leq 1.0$, is governed by the Fresnel function

$$f(\alpha_i, r) = \frac{1}{2}\frac{\sin^2(\alpha_i - \alpha_t)}{\sin^2(\alpha_i + \alpha_t)}\left[1 + \frac{\cos^2(\alpha_i + \alpha_t)}{\cos^2(\alpha_i - \alpha_t)}\right] \tag{5}$$

The transmission angle α_t of light into the dielectric surface is given by Snell's law $r = (\sin\alpha_i)/(\sin\alpha_t)$. The parameter n is the index of refraction of the dielectric medium. When light is transmitted from air into a dielectric $r = n$ and $\alpha_i = \theta_i$. However, when transmission is from a dielectric into air, then $r = 1/n$ and $\alpha_i = \sin^{-1}[(\sin\theta_s)/n]$. The Wolff model deviates from the Lambertian form, i.e. $\cos\theta_i$, when the Fresnel terms become significant. Almost all commonly found dielectric materials have an index of refraction, n, in the range $[1.4, 2.0]$. As a result the Fresnel function is weakly dependent upon the index of refraction for most dielectrics. The value of the scaling factor ϱ is very nearly constant over most illumination conditions [15].

Our approach is to replace the geometrical term F^2 in the B-K model with a Fresnel correction term. Whereas Vernold and Harvey [14] have replaced the F^2 term by $\cos\theta_i$, we replace it by the factor $\cos\theta_i$ multiplied by two Fresnel terms (for incidence and for reflection). The correction term is

$$F^2_{FC2}(\theta_i, \theta_s, n) = [1 - f(\theta_i, n)]\{1 - f(\sin^{-1}[(\sin\theta_s)/n], 1/n)\}\cos\theta_i \tag{6}$$

We also investigate the behavior of an alternative Fresnel correction term in which $F^2_{FC1} = F^2_{FC2}\cos\theta_i$.

4 Experiments

We have fitted the various variants of the B-K model to the CUReT data using a least-squares fitting technique, using both exponential (denoted X-E) and Gaussian (denoted X-G) correlation functions. To do this we seek the slope parameter values that minimize the mean-squared-error

$$MSE = \sum_{k=1}^{K} [\mathcal{L}_o^M(\theta_i^k, \phi_i^k, \theta_s^k, \phi_s^k) - \mathcal{L}_o^D(\theta_i^k, \phi_i^k, \theta_s^k, \phi_s^k)]^2 \tag{7}$$

where $\mathcal{L}_o^M$ is the normalized radiance value predicted by the model, $\mathcal{L}_o^D$ that obtained from the BRDF data, and k runs over the index number of the illumination configurations used (there are $K = 205$ configurations in the CUReT database). To locate the least-squares (LSE) slope parameter values we test 200 equally spaced values in the interval [0.1, 1.1] for the model variants with a Gaussian correlation function and in the interval [0.5, 2.5] for those with an

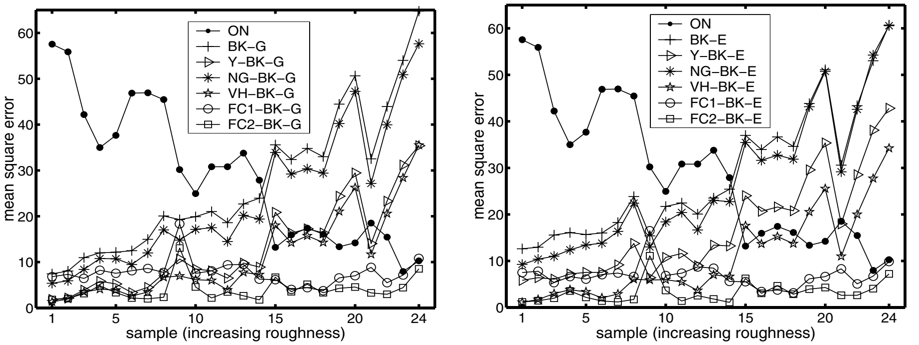
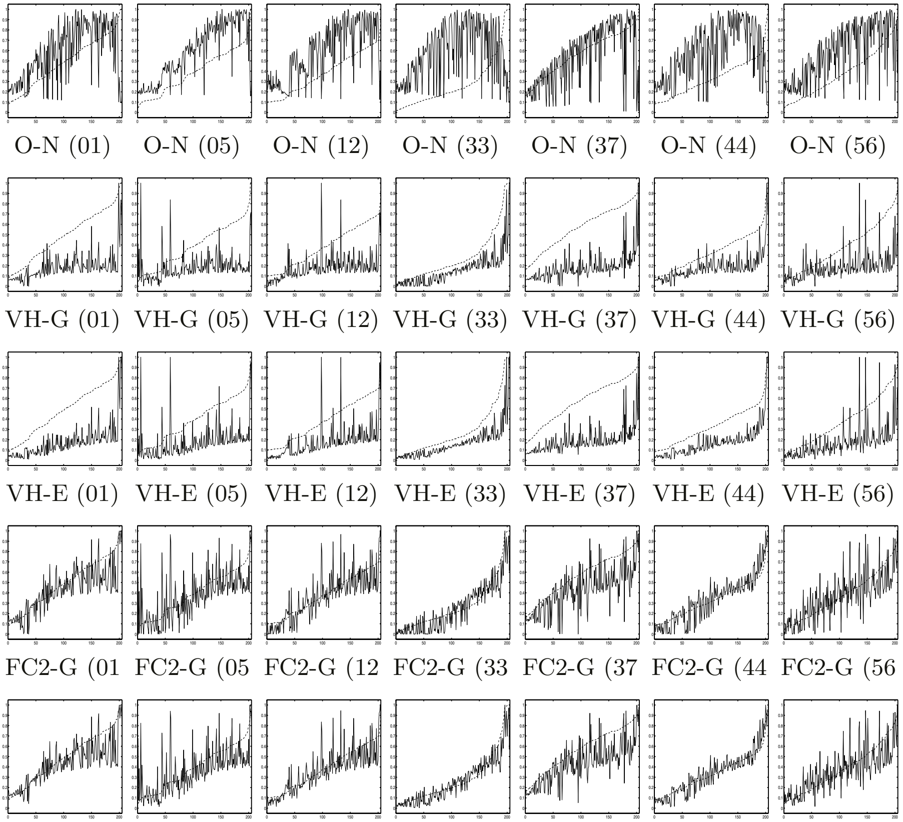


Fig. 1. Least-squared-error for the Oren-Nayar model and the B-K model variants with the Gaussian (left) and exponential (right) correlation functions versus sample index sorted according to increasing slope estimates (corresponding to the FC2-E variant).

exponential correlation function. In Fig. 1 we summarize the results. Here we have sorted the samples according to increasing slope parameter using the FC2-E variant, and have plotted the least-squared-error (LSE) between the model and the data as a function of surface sample number. The left-hand plot is for the Gaussian correlation function and the right-hand plot is for the exponential correlation function. There are a number of features that drawn from the plot. First, the LSE for the Oren-Nayar model (ON) decreases with increasing roughness, i.e. the fit of the model to the different data samples improves with increasing roughness. For both the original B-K model (BK-G and BK-E) and the Nieto-Garcia model (NG-BK-G and NG-BK-E), on the other hand, the LSE increases with increasing roughness. In other words, the models work best for slightly-rough surfaces. A similar, but less marked effect is exhibited by the V-H modification of the B-K model. For both Fresnel correction variants of the B-K model, in particular FC2, the LSE is consistently lower than all of the alternatives, it is also relatively insensitive to the roughness order. Hence, the model is a better fit to the data over all scales of roughness.

There are a number of conclusions that can be drawn from these plots. First we consider the behavior of the Oren-Nayar model. From Fig. 1 it is clear that the model gives the best results for very-rough surface samples. Next we turn our attention to the original B-K model and the modification to it by Vernold and Harvey. The V-H modification (VH-E and VH-G) gives LSE values that are always lower than those obtained with the original model (BK-E and BK-G). For the modified B-K model it is difficult to distinguish between the exponential (VH-E) and Gaussian (VH-G) correlation functions on the basis of LSE. For the original B-K model, the Gaussian correlation function (BK-G) always gives a slightly lower LSE than the exponential one (BK-E). Finally, when Ogilvy's expression for surfaces with a Gaussian correlation function (Y-BK-G) is used then the results for certain surface samples are better than those obtained using the original B-K model. These conclusions are supported by Fig. 2. In their study of this data, Koenderink et al. [6] have noted that the observations are



O-N (01) O-N (05) O-N (12) O-N (33) O-N (37) O-N (44) O-N (56)

VH-G (01) VH-G (05) VH-G (12) VH-G (33) VH-G (37) VH-G (44) VH-G (56)

VH-E (01) VH-E (05) VH-E (12) VH-E (33) VH-E (37) VH-E (44) VH-E (56)

FC2-G (01 FC2-G (05 FC2-G (12 FC2-G (33 FC2-G (37 FC2-G (44 FC2-G (56

FC2-E (01 FC2-E (05 FC2-E (12 FC2-E (33 FC2-E (37 FC2-E (44 FC2-E (56

Fig. 2. Selected plots for the surface samples 1, 5, 12, 33, 37, 44 and 56: normalized radiance versus data index (0-205) corresponding to increasing radiance data: model predictions (solid curves, Table 1) and data measurements (dashed curves).

relatively sparse, and hence their visualization over the four angular variables is not a straightforward task. To overcome these problems, we have sorted the BRDF measurement indices in the order of increasing radiance, so that we can compare them with different reflectance models more conveniently. This method of sorting results in slowly varying data curves, but rather irregular model curves.

We have shown only the plots corresponding to the Oren-Nayar model together with those for the VH-G, VH-E, FC2-G and FC2-E model variants. These variants of the B-K model result in the best fits with the data when compared to the remaining alternatives. Here, for all surface samples studied, the FC2-E and FC2-G models give the qualitatively better fits. The V-H modification gives rather poor results. The Oren-Nayar model performs best on the very-rough samples. From the corresponding plots, the fits appear qualitatively good, except perhaps at large radiance values. However, for the smooth and moderately rough samples, the Oren-Nayar model does not perform well when compared with the alternatives.

5 Conclusions

In this paper we have explored how to extend the Beckmann-Kirchhoff (B-K) theory [1] to surfaces of intermediate roughness. The most successful existing version of the B-K model is the modification due to Vernold-Harvey [14]. This modification is aimed at improving the B-K model for large incidence and scattering angles. To develop a modification of the model which can be used for moderately-rough surfaces, we have exploited the Fresnel coefficient in a manner similar to that of Wolff [15]. This allows us to incorporate the effects of refractive attenuation into the B-K theory.

References

1. P. Beckmann and A. Spizzichino, *The Scattering of Electromagnetic Waves from Rough Surfaces*, Pergamon, New York, 1963.
2. J.M. Bennett, and L. Mattsson, *Introduction to Surface Roughness and Scattering: 2nd ed.*, Optical Society of America, Washington, D.C., 1999.
3. CUReT database, www.cs.columbia.edu/CAVE/curet.
4. J.E. Harvey, C.L. Vernold, A. Krywonos and P.L. Thompson, "Diffracted Radiance: A Fundamental Quantity in a Non-paraxial Scalar Diffraction Theory," *Applied Optics*, vol. 38, no. 31, 1999, pp. 6469-6481.
5. X.D. He, K.E. Torrance, F.X. Sillion and D.P. Greenberg, "A Comprehensive Physical Model for Light Reflection," *ACM Computer Graphics*, vol. 25, 1991, pp. 175-186.
6. J.J. Koenderink, A.J. van Doorn and M. Stavridi, "Bidirectional Reflection Distribution Function expressed in terms of Surface Scattering Modes," *European Conference on Computer Vision*, vol. 2, 1996, pp. 28-39.
7. S.K. Nayar, K. Ikeuchi and T. Kanade, "Surface Reflection: Physical and Geometrical Perspectives," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13, no. 7, 1991, pp. 611-634.
8. M. Nieto-Vesperinas and N. Garcia, "A Detailed Study of the Scattering of Scalar Waves from Random Rough Surfaces," *Optica Acta*, vol. 28, no. 12, 1981, pp. 1651-1672.
9. J.A. Ogilvy, *Theory of Wave Scattering from Random Rough Surfaces*, Adam Hilger, Bristol, 1991.
10. M. Oren and S.K. Nayar, "Generalization of the Lambertian Model and Implications for Machine Vision," *International Journal on Computer Vision*, vol. 14, no. 3, 1995, pp. 227-251.
11. B.T. Phong, "Illumination for Computer Generated Pictures," *Communications ACM*, vol. 18, 1975, pp. 311-317.
12. K.E. Torrance and E.M. Sparrow, "Theory for Off-Specular Reflection from Roughened Surfaces," *Journal of the Optical Society of America*, vol. 57, no. 9, 1967, pp. 1105-1114.
13. B. van Ginneken, M. Stavridi and J. Koenderink, "Diffuse and Specular Reflectance from Rough Surfaces," *Applied Optics*, vol. 37, no. 1, 1998, pp. 130-139.
14. C.L. Vernold, and J.E. Harvey, "A Modified Beckmann-Kirchoff Scattering Theory for Non-paraxial Angles,"*Proceedings of the SPIE,* vol. 3426, 1998, pp. 51-56.
15. L.B. Wolff "Diffuse Reflectance Model for Smooth Dielectric Surfaces," *Journal of the Optical Society of America A*, vol. 11, no. 11, 1994, pp. 2956-2968.

Multi-scale Cortical Keypoint Representation for Attention and Object Detection

João Rodrigues[1] and Hans du Buf[2]

[1] University of Algarve, Escola Superior Tecnologia, Faro, Portugal
[2] University of Algarve, Vision Laboratory, FCT, Faro, Portugal

Abstract. Keypoints (junctions) provide important information for focus-of-attention (FoA) and object categorization/recognition. In this paper we analyze the multi-scale keypoint representation, obtained by applying a linear and quasi-continuous scaling to an optimized model of cortical end-stopped cells, in order to study its importance and possibilities for developing a visual, cortical architecture. We show that keypoints, especially those which are stable over larger scale intervals, can provide a hierarchically structured saliency map for FoA and object recognition. In addition, the application of non-classical receptive field inhibition to keypoint detection allows to distinguish contour keypoints from texture (surface) keypoints.

1 Introduction

Models of cells in the visual cortex, i.e. simple, complex and end-stopped, have been developed, e.g. [4]. In addition, several inhibition models [3, 11], keypoint detection [4, 13, 15] and line/edge detection schemes [3, 13, 14], including disparity models [2, 9, 12], have become available. On the basis of these models and processing schemes, it is now possible to create a cortical architecture for figure-background separation [5, 6] and visual attention or focus-of-attention (FoA), bottom-up or top-down [1, 10], and even for object categorization and recognition.

In this paper we will focus on keypoints, for which Heitger et al. [4] developed a single-scale basis model of single and double end-stopped cells. Würtz and Lourens [15] presented a multi-scale approach: spatial stabilization is obtained by averaging keypoint positions over a few neighboring micro-scales. We [13] also applied multi-scale stabilization, but focused on integrating line/edge, keypoint and disparity detection, including the classification of keypoint structure (e.g. T, L, K junctions). Although the approaches in [13, 15] were multi-scale, the aim was stabilization at one (fine) scale. Here we will go into a truly multi-scale analysis: we will analyze the multi-scale keypoint representation, from very fine to very coarse scales, in order to study its importance and possibilities for developing a cortical architecture, with an emphasis on FoA. In addition, we will include a new aspect, i.e. the application of non-classical receptive field (NCRF) inhibition [3] to keypoint detection, in order to distinguish between object structure and surface textures.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 255–262, 2005.

2 End-Stopped Models and NCRF Inhibition

Gabor quadrature filters provide a model of cortical simple cells [8]. In the spatial domain (x,y) they consist of a real cosine and an imaginary sine, both with a Gaussian envelope. A receptive field (RF) is denoted by (see e.g. [3]):

$$g_{\lambda,\sigma,\theta,\varphi}(x,y) = \exp\left(-\frac{\tilde{x}^2+\gamma\tilde{y}^2}{2\sigma^2}\right)\cdot\cos(2\pi\frac{\tilde{x}}{\lambda}+\varphi),$$

$$\tilde{x} = x\cos\theta + y\sin\theta; \tilde{y} = y\cos\theta - x\sin\theta,$$

where the aspect ratio $\gamma = 0.5$ and σ determines the size of the RF. The spatial frequency is $1/\lambda$, λ being the wavelength. For the bandwidth σ/λ we use 0.56, which yields a half-response width of one octave. The angle θ determines the orientation (we use 8 orientations), and φ the symmetry (0 or $\pi/2$). We apply a linear scaling between $f_{\min}$ and $f_{\max}$ with, at the moment, hundreds of contiguous scales.

The responses of even and odd simple cells, which correspond to the real and imaginary parts of a Gabor filter, are obtained by the convolution of the input image with the RF, and are denoted by $R^E_{s,i}(x,y)$ and $R^O_{s,i}(x,y)$, s being the scale and i the orientation ($\theta_i = i\pi/(N_\theta - 1)$) and N_θ the number of orientations. In order to simplify the notation, and because the same processing is done at all scales, we drop the subscript s. The responses of complex cells are modelled by the modulus

$$C_i(x,y) = [\{R^E_i(x,y)\}^2 + \{R^O_i(x,y)\}^2]^{1/2}.$$

There are two types of end-stopped cells [4, 15], i.e. single (S) and double (D). If $[\cdot]^+$ denotes the suppression of negative values, and $\mathcal{C}_i = \cos\theta_i$ and $\mathcal{S}_i = \sin\theta_i$, then

$$S_i(x,y) = [C_i(x+d\mathcal{S}_i, y-d\mathcal{C}_i) - C_i(x-d\mathcal{S}_i, y+d\mathcal{C}_i)]^+ ;$$

$$D_i(x,y) = \left[C_i(x,y) - \frac{1}{2}C_i(x+2d\mathcal{S}_i, y-2d\mathcal{C}_i) - \frac{1}{2}C_i(x-2d\mathcal{S}_i, y+2d\mathcal{C}_i)\right]^+ .$$

The distance d is scaled linearly with the filter scale s, i.e. $d = 0.6s$. All end-stopped responses along straight lines and edges need to be suppressed, for which we use tangential (T) and radial (R) inhibition:

$$I^T(x,y) = \sum_{i=0}^{2N_\theta-1} [-C_{i \bmod N_\theta}(x,y) + C_{i \bmod N_\theta}(x+d\mathcal{C}_i, y+d\mathcal{S}_i)]^+ ;$$

$$I^R(x,y) = \sum_{i=0}^{2N_\theta-1} \left[C_{i \bmod N_\theta}(x,y) - 4\cdot C_{(i+N_\theta/2) \bmod N_\theta}(x+\frac{d}{2}\mathcal{C}_i, y+\frac{d}{2}\mathcal{S}_i)\right]^+ ,$$

where $(i+N_\theta/2) \bmod N_\theta \perp i \bmod N_\theta$.

The model of non-classical receptive field (NCRF) inhibition is explained in more detail in [3]. We will use two types: (a) anisotropic, in which only responses obtained for the same preferred RF orientation contribute to the suppression, and (b) isotropic, in which all responses over all orientations equally contribute to the suppression.

The anisotropic NCRF (A-NCRF) model is computed by an inhibition term $t^A_{s,\sigma,i}$ for each orientation i, as a convolution of the complex cell response C_i with the weighting function w_σ, with $w_\sigma(x,y) = [DoG_\sigma(x,y)]^+/\|[DoG_\sigma]^+\|_1$, $\|\cdot\|_1$ being the L_1 norm, and

$$DoG_\sigma(x,y) = \frac{1}{2\pi(4\sigma)^2}\exp(-\frac{x^2+y^2}{2(4\sigma)^2}) - \frac{1}{2\pi\sigma^2}\exp(-\frac{x^2+y^2}{2\sigma^2}).$$

The operator $b^A_{s,\sigma,i}$ corresponds to the inhibition of $C_{s,i}$, i.e. $b^A_{s,\sigma,i} = [C_{s,i} - \alpha t^A_{s,\sigma,i}]^+$, with α controlling the strength of the inhibition.

The isotropic NCRF (I-NCRF) model is obtained by computing the inhibition term $t^I_{s,\sigma}$ which does not dependent on orientation i. For this we construct the maximum response map of the complex cells $\tilde{C}_s = \max\{C_{s,i}\}$, with $i = 0, ...N_\theta - 1$. The isotropic inhibition term $t^I_{s,\sigma}$ is computed as a convolution of the maximum response map $\tilde{C}_s$ with the weighting function w_σ, and the isotropic operator is $b^I_{s,\sigma} = [\tilde{C}_s - \alpha t^I_{s,\sigma}]^+$.

3 Keypoint Detection with NCRF Inhibition

NCRF inhibition permits to suppress keypoints which are due to texture, i.e. textured parts of an object surface. We experimented with the two types of NCRF inhibition introduced above, but here we only present the best results which were obtained by I-NCRF at the finest scale.

All responses of the end-stopped cells $S(x,y) = \sum_{i=0}^{N_\theta-1} S_i(x,y)$ and $D(x,y) = \sum_{i=0}^{N_\theta-1} D_i(x,y)$ are inhibited in relation to the complex cells (by $b^I_{s,\sigma}$), i.e. we use $\alpha = 1$, and obtain the responses $\tilde{S}$ and $\tilde{D}$ of S and D that are above a small threshold of $b^I_{s,\sigma}$. Then we apply $I = I^T + I^R$ for obtaining the keypoint maps $K^S(x,y) = \tilde{S}(x,y) - gI(x,y)$ and $K^D(x,y) = \tilde{D}(x,y) - gI(x,y)$, with $g \approx 1.0$, and then the final keypoint map $K(x,y) = \max\{K^S(x,y),\, K^D(x,y)\}$.

Figure 1 presents, from left to right, input images and keypoints detected (single scale), before and after I-NCRF inhibition. The top image shows part of a building in Estoril ("Castle"). The middle images show two leaves, and the bottom one is a traffic sign (also showing, to the right, vertex classification with micro-scale stability, see [13]). Most important keypoints have been detected, and after inhibition contour-related ones remain. Almost all texture keypoints have been suppressed, although some remain (Castle image) because of strong local contrast and the difficulty of selecting a good threshold value without eliminating important contour keypoints (see Discussion).

4 Multi-scale Keypoint Representation

Here we focus on the multi-scale representation. Although NCRF inhibition can be applied at each scale, we will not do this for two reasons: (a) we want to

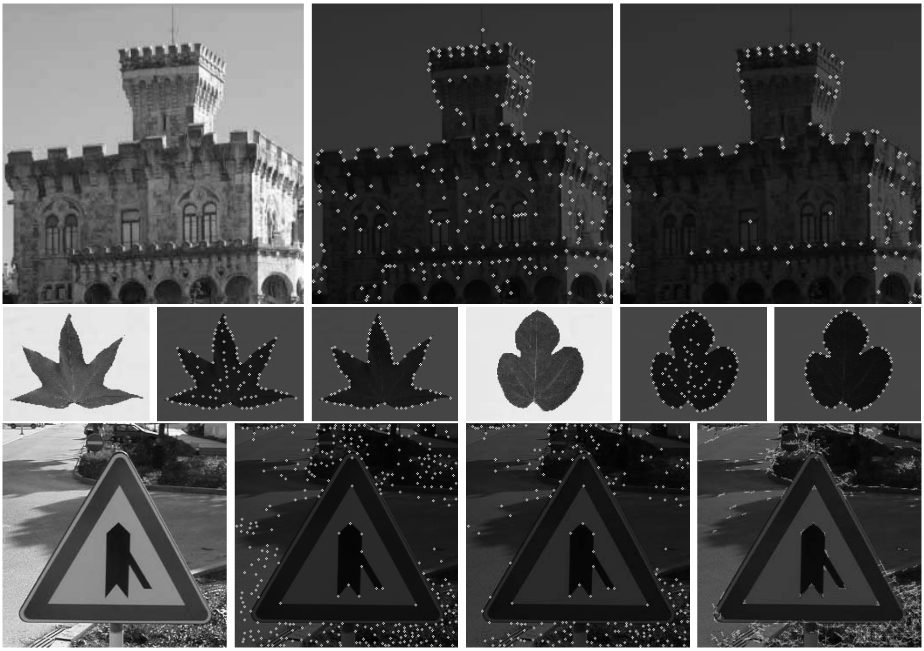


Fig. 1. Keypoints without and with NCRF inhibition; see text.

study keypoint behavior in scale space for applications like FoA, and (b) in many cases a coarser scale, i.e. increased RF size, will automatically eliminate detail (texture) keypoints. In the multi-scale case keypoints are detected the same way as done above, but now by using $K_s^S(x,y) = S_s(x,y) - gI_s(x,y)$ and $K_s^D(x,y) = D_s(x,y) - gI_s(x,y)$.

For analyzing keypoint stability we create an almost continuous, linear, scale space. In the case of Fig. 2, which shows the (projected) trajectories of detected keypoints over scale in the case of a square and a star, we applied 288 scales with $4 \leq \lambda \leq 40$. Figure 2 illustrates the general behavior: at small scales contour keypoints are detected, at coarser scales their trajectories converge, and at very coarse scales there is only one keypoint left near the center of the object. However, it also can be seen (star object) that there are scale intervals where keypoints are unstable, even scales at which keypoints disappear and other scales at which they appear. (Dis)appearing keypoints are due to the size of the RFs in relation to the structure of the objects, in analogy with Gaussian scale space [7]. Unstable keypoints can be eliminated by (a) requiring stability over a few neighboring micro-scales [13], i.e. keep keypoints that do not change position over 5 scales, the center one and two above and two below (Fig. 2e), or (b) requiring stability over at least N_s neighboring scales (Fig. 2f and 2g with $N_s = 10$ and 40, respectively).

The leftmost five columns in Fig. 3 illustrate that similar results are obtained after blurring, adding noise, rotation and scaling of an object (a leaf), whereas

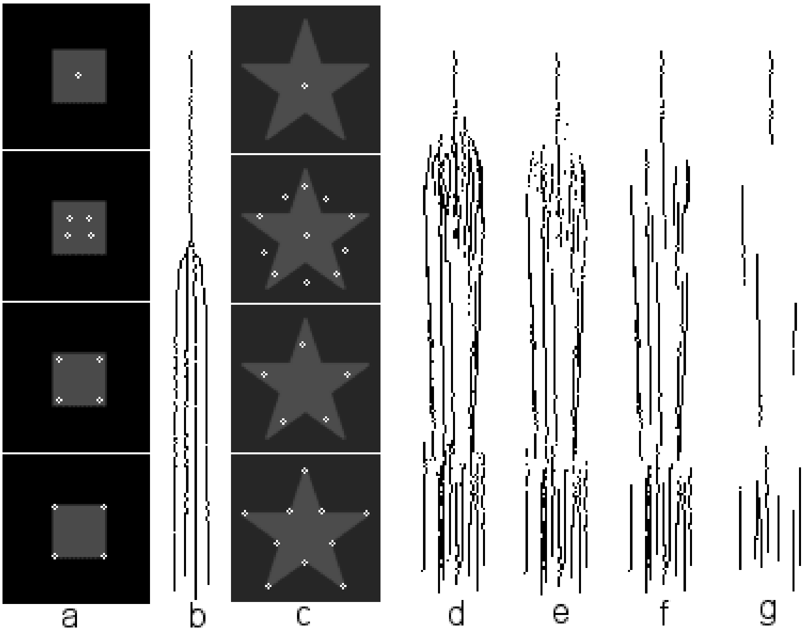


Fig. 2. Keypoint scale space, with finest scale at the bottom. From left to right: (a) square; (b) projected 3D keypoint trajectories of square; (c) and (d) star and projected trajectories; (e) micro-scale stability; (f) and (g) stability over at least 10 and 40 scales respectively.

the last two columns show results for other leave shapes. In all cases, important contour keypoints remain at medium scales and texture keypoints disappear, without applying NCRF inhibition.

With respect to object categorization/recognition, a coarse-to-fine scale strategy appears to be feasible. Figure 4 shows an image with four objects, i.e. two leaves, a star and a van from a traffic sign (see [13]). At very coarse scales the keypoints indicate centers of objects. In the case of the elongated van, an even coarser scale is required. Going from coarse to fine scales, keypoints will indicate more and more detail, until the finest scale at which essential landmarks on contours remain. In reality, the keypoint information shown will be completed by line/edge and disparity (3D) information.

Figure 5 shows that a coarse-to-fine strategy is also feasible in the case of real scenes, i.e. the tower of the Castle image. At coarse scales keypoints indicate the shape of the tower; at finer scales appear structures like the battlements, whereas the corners of the battlements appear at the finest scales. Here we did not apply NCRF inhibition to all scales in order to show that the multi-scale approach selectively "sieves" according to structure detail and contrast.

Another element of an object detection scheme is focus-of-attention by means of a saliency map, i.e. the possibility to inspect, serially or in parallel, the most important parts of objects or scenes. If we assume that retinotopic projection is

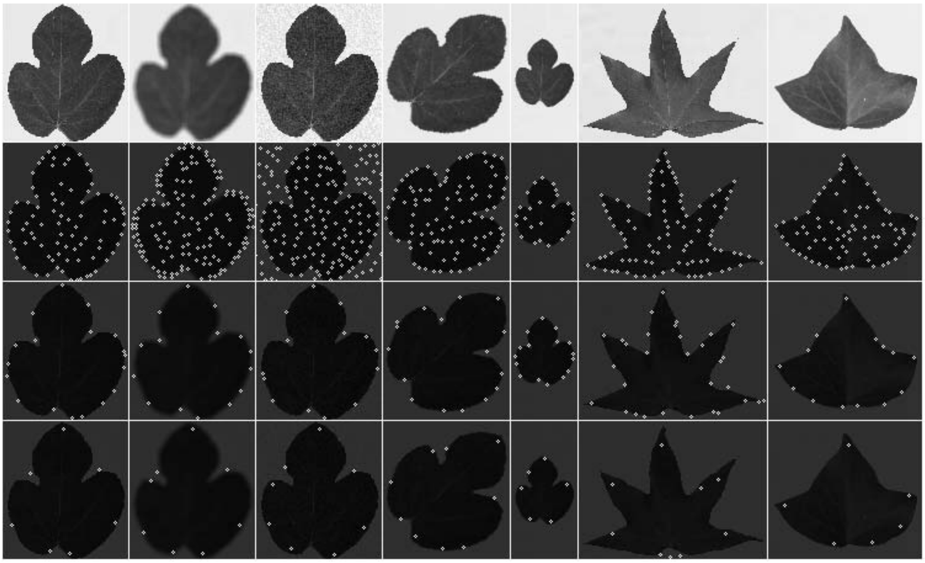


Fig. 3. From left to right: ideal image, blurred, with added noise, rotated and scaled leaf, plus two other leaves. From fine (2nd line) to medium scale (bottom line).

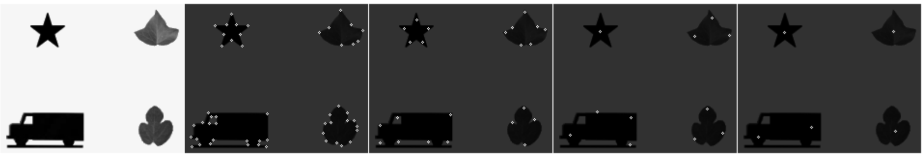


Fig. 4. Object detection from coarse (right) to fine (left) scales ($4 \leq \lambda \leq 50$).

maintained throughout the visual cortex, the activity of keypoint cells at position (x, y) can be easily summed over scale s. At the positions where keypoints are stable over many scales, this summation map, which could replace or contribute to a saliency map [10], will show distinct peaks at centers of objects, important structures and contour landmarks. The height of the peaks can provide information about the relative importance. This is shown in Fig. 6. In addition, this summation map, with some simple processing of the projected trajectories of unstable keypoints, like lowpass filtering and non-maximum suppression, might solve the segmentation problem: the object center is linked to important structures, and these are linked to contour landmarks. Such a data stream is data-driven and bottom-up, and could be combined with top-down processing from inferior temporal cortex (IT) in order to actively probe the presence of objects in the visual field [1]. In addition, the summation map with links between the peaks might be available at higher cortical levels, where serial processing occurs for e.g. visual search.

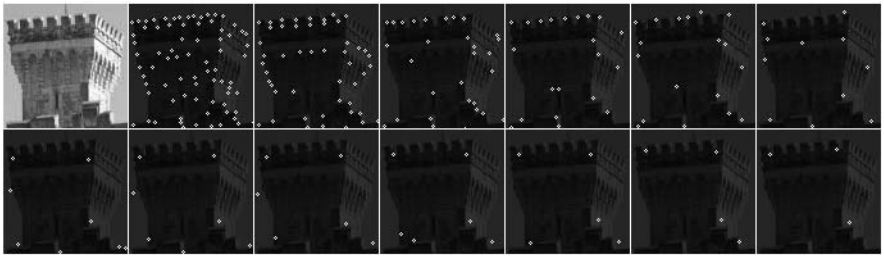


Fig. 5. Keypoint scale-space without NCRF inhibition. From left to right and top to bottom increasing scale ($4 \leq \lambda \leq 50$).

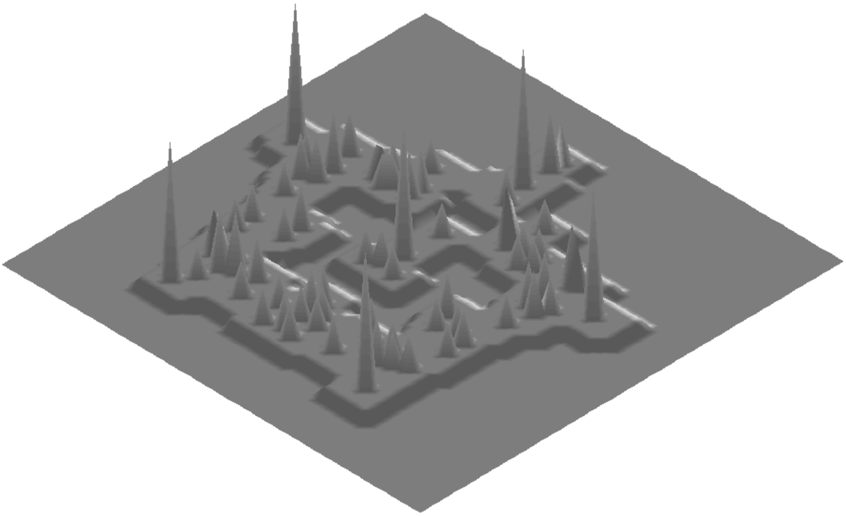


Fig. 6. 3D visualization of the keypoint summation map of the star.

5 Conclusions

The primary visual cortex contains low-level processing "engines" for retinotopic feature extraction. These include multi-scale lines and edges, bars and gratings, disparity and keypoints. Mainly being data-driven, these engines feed higher processing levels, for example for translation, rotation and scale invariant object representations, visual attention and search, until object recognition.

To the best of our knowledge, this is the first study to analyze the importance of multi-scale keypoint representation for e.g. focus-of-attention and object recognition. We showed that the trajectories of keypoints in scale space may be quite complex, but also that keypoints are stable at important structures. In general, at coarse scales keypoints can be expected at centers of objects, at finer scales at important structures, until they cover finest details. We also showed that retinotopic summation of "keypoint-cell activity" over scale provides very useful information for a saliency map (FoA), and even could solve the segmentation problem by bounding objects and linking structures within objects. It seems that the multi-scale keypoint representation, obtained by a linear scaling of cortical end-stopped operators, might be the most important component in building a complete cortical architecture. However, much more information is available through line and edge cells, bar and grating cells, and disparity-tuned cells. In addition, data-driven and bottom-up signals must be used together with top-down or feedback signals coming from higher processing levels.

Finally, it should be mentioned that the hundreds of quasi-continuous scales used here, which is computationally very expensive, can be seen as an abstraction of cortical reality: in reality, there may be an octave or half-octave RF organization, with at each level adaptivity (plasticity) in order to stabilize detection results. Such a scheme, and its application to e.g. FoA, has not yet been explored.

References

1. G. Deco and E.T. Rolls. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Res.*, (44):621–642, 2004.
2. D.J. Fleet, A.D. Jepson, and M.R.M. Jenkin. Phase-based disparity measurement. *CVGIP: Image Understanding*, 53(2):198–210, 1991.
3. C. Grigorescu, N. Petkov, and M.A. Westenberg. Contour detection based on nonclassical receptive field inhibition. *IEEE Tr. Im. Proc.*, 12(7):729–739, 2003.
4. F. Heitger et al. Simulation of neural contour mechanisms: from simple to end-stopped cells. *Vision Res.*, 32(5):963–981, 1992.
5. J.M. Hupe et al. Cortical feedback improves discrimination between figure and background by v1, v2 and v3 neurons. *Nature*, 394(6695):784–787, 1998.
6. J.M. Hupe et al. Feedback connections act on the early part of the responses in monkey visual cortex. *J. Neurophysiol.*, 85(1):134–144, 2001.
7. J.J. Koenderink. The structure of images. *Biol. Cybern.*, 50(5):363–370, 1984.
8. T.S. Lee. Image representation using 2D Gabor wavelets. *IEEE Tr. PAMI*, 18(10):pp. 13, 1996.
9. I. Ohzawa, G.C. DeAngelis, and R.D. Freeman. Encoding of binocular disparity by complex cells in the cat's visual cortex. *J. Neurophysiol.*, 18(77):2879–2909, 1997.
10. D. Parkhurst, K. Law, and E. Niebur. Modelling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123, 2002.
11. N. Petkov, T. Lourens, and P. Kruizinga. Lateral inhibition in cortical filters. *Proc. Int. Conf. Dig. Sig. Proc. and Inter. Conf. on Comp. Appl. to Eng. Sys.*, Nicosa, Cyprus:122–129, July 14-16 1993.
12. J. Rodrigues and J.M.H. du Buf. Vision frontend with a new disparity model. *Early Cognitive Vision Workshop, Isle of Skye, Scotland*, 28 May - 1 June 2004.
13. J. Rodrigues and J.M.H. du Buf. Visual cortex frontend: integrating lines, edges, keypoints and disparity. *Proc. Int. Conf. Image Anal. Recogn.(ICIAR)*, Springer LNCS 3211(1):664–671, 2004.
14. J.H. van Deemter and J.M.H. du Buf. Simultaneous detection of lines and edges using compound Gabor filters. *Int. J. Patt. Rec. Artif. Intell.*, 14(6):757–777, 1996.
15. R.P. Würtz and T. Lourens. Corner detection in color images by multiscale combination of end-stopped cortical cells. *Image and vision computing*, 18(6-7):531–541, 2000.

Evaluation of Distances Between Color Image Segmentations

Jaume Vergés-Llahí and Alberto Sanfeliu

Institut de Robòtica i Informàtica Industrial
Technological Park of Barcelona, U Building
{jverges,asanfeliu}@iri.upc.es

Abstract. We illustrate the problem of comparing images by means of their color segmentations. A group of seven distances are proposed within the frame of the Integrated Region Matching distance and the employ of Multivariate Gaussian Distributions (MGD) for the color description of image regions. The performance of these distances is examined in tasks such as image retrieval and object recognition using the two segmentation algorithms in [1] and [2]. The best overall results are obtained for both tasks using the graph–partition approach along with the Fréchet distance, outperforming other distances in comparing MGDs.

Keywords: color segmentation, image retrieval, object identification.

1 Introduction

The aim of this paper consists in comparing images on the base of their color segmentation. The necessity for this arises from well–known tasks such as object recognition, image indexing and retrieval, as well as others related to mobile robotics. Content–based comparison of image segmentations can be viewed as an object recognition problem. Nevertheless, our situation is slightly opener than that of classical object recognition since segmentation of real images is most of the time imperfect. We need a more flexible way to compare segmented images on the basis of their content whether segmentation is perfect or not.

Consequently, our concern in this paper is to study different distances between images based on their color segmentation which can cope with imperfect segmentations. Such a measure would be helpful both in finding images in a database, identifying objects and comparing image by their content. Our aims heed the set of techniques known as *Content–Based Image Retrieval* (CBIR) since those methods develop measures among images based on their content to effectively indexing and searching in large–scale image databases. More precisely, CBIR is the set of techniques for retrieving semantically–relevant images from an image database on automatically–derived image features.

The reasons for using this framework are mainly two. First, these techniques are not focused on finding an exact identification, rather more flexible retrieval schemes are usually undergone. Second, they tend to cope with imperfect or inexact segmentations. Therefore, this paper proposes some distances between regions in the context of a CBIR distance between segmentations to evaluate their performance in tasks such as object identification and image retrieval.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 263–270, 2005.

2 Related Work

Because of the lack of space, we only refer to some of the best–known CBIR techniques. CBIR for general purpose databases is still a great challenging problem because of the size of such databases, the difficulties in both understanding images and formulating a query, and the properly evaluation of results. A common problem for all CBIR systems is to extract a *signature* from images based on their pixel values and to define an effective rule for comparing images.

The signature, whose components are called *features*, serves as the image representation. The important reason for using signatures, besides the significant compression of the image representation, is that of improving the correlation between image pixels and image semantics, i.e., understanding the image content by means of its pixel values. Most existing general–purpose CBIR systems roughly fall into the next three categories, depending on the image signatures

– *Histograms:* IBM QBIC [3], MIT Photobook [4].
– *Color layouts:* VIRAGE [5], Columbia VisualSEEK and WebSEEK [6], Stanford WBIIS [7], and WALRUS [8].
– *Region–based systems:* NeTra [9], Blobworld [10], SIMPLIcity [11], SNL [12].

After extracting signatures, the next step is to determine a *comparison rule*, which includes a *querying scheme* – global or partial search – and the definition of a *similarity measure* between two images. Despite numerous types of signatures can be employed, such as shape and texture, our concern here is just those based on the color information of image pixels.

3 Distance Between Segmented Images

Intuitively, a region–based distance between two segmented images can be defined as the total amount of differences between corresponding regions. It is clear that only regions which are likely the same in both segmentations must be taken into account. Otherwise, the measure would be biased. In addition, any distance based on image segmentation should be tolerant to inaccuracies. To define this kind of measures, an attempt to match regions must be carried out somehow, despite it could be error prone and time consuming. Nevertheless, matching has to be softened by allowing one region of an image to be matched to several regions in another one since segmentations are not perfect.

Furthermore, a similarity measure among regions is equivalent to a distance between sets of points in a feature space. Every point in the space corresponds to a descriptor of a region. Although the distance between two points in a feature space can be chosen from a variety, it is not obvious how to define a distance between two groups or distributions of points. In this paper, we first describe the region descriptor applied throughout the work to account for color distributions. Afterwards, a number of different approaches to measure the similarity between these descriptors are proposed. Finally, we describe how to group all these measures into only one distance between two segmented images.

3.1 Multivariate Gaussian Distributions

We describe region colors as a *Multivariate Gaussian Distribution* (MGD) of probability for its simplicity and compactness, along with its good mathematical properties. Besides, it is a natural way of introducing *Gaussian Mixture Models* (GMM) as a method of representing segmented images [2]. Hence, let the color of a region be a random variable $X \sim N_d(\mu, \Sigma)$ distributed as

$$P(\mathbf{x}|\boldsymbol{\Theta}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp\left(-\frac{1}{2}(\mathbf{x}-\mu)\,\Sigma^{-1}(\mathbf{x}-\mu)^t\right) \qquad (1)$$

where $\mathbf{x} \in \mathbb{R}^d$ is a sample point, $\boldsymbol{\Theta} = \{\mu, \Sigma\}$ are the mean and covariance matrix of P, respectively, and d is the space dimension[1]. We note as $\mathcal{X}$ the distribution of the random variable X. So, for each region $\mathcal{R}_i$ in image $\mathcal{I}$ there will be a distribution $\mathcal{X}_i$ whose parameters are $\{\overline{\mathbf{x}}_i, \boldsymbol{\Sigma}_i, \omega_i\}$, being ω_i the region weight, and $\overline{\mathbf{x}}_i$ and $\boldsymbol{\Sigma}_i$ are the sample mean and covariance matrix, respectively.

3.2 Distances Between Image Regions

Now, we discuss seven diverse distances between two regions $\mathcal{R}_x$ and $\mathcal{R}_y$, expressed in terms of their distributions $\mathcal{X}$ and $\mathcal{Y}$. Nevertheless, this is not the discriminant case where a distance $\mathcal{D}(\mathbf{x}, \mathcal{Y})$ between a point $\mathbf{x}$ and a distribution $\mathcal{Y}$ is computed to know whether $\mathbf{x}$ belongs to $\mathcal{Y}$. Rather, we need to estimate the distance between two whole MGDs, i.e., $\mathcal{D}(\mathcal{X}, \mathcal{Y})$, corresponding to random variables $X \sim N_d(\mu_x, \Sigma_x)$ and $Y \sim N_d(\mu_y, \Sigma_y)$.

Euclidean Distance. It consists only in computing the Euclidean distance between the two means, $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$, each of them representing a distribution center

$$\mathcal{D}^2(\mathcal{X}, \mathcal{Y}) = \|\overline{\mathbf{x}} - \overline{\mathbf{y}}\|^2 \qquad (2)$$

Its simplicity is both a pro and a con, since it neither appraises the shape of the distributions nor its relative size, conveyed in the covariance matrices.

Mahalanobis Distance. Shape and orientation could be interesting when trying to compare distributions, for example, whenever two distributions are centered at the same point. Mahalanobis distance introduces these subtleties into account by using as a metric the inverse of the covariance matrix of the particular distribution. This way, the squared distance between a sample point $\mathbf{y}$ and a distribution $\mathcal{X}$ is computed as $\mathcal{D}^2(\mathbf{y}, \mathcal{X}) = (\mathbf{y} - \overline{\mathbf{x}})\,\boldsymbol{\Sigma}_x^{-1}(\mathbf{y} - \overline{\mathbf{x}})^t$.

Since these distances are bilinear, it is true that the mean distance to a distribution of a set of points $\{\mathbf{y}_i\}_{i=1,\ldots,n_y}$ equals to the distance between the sample mean $\overline{\mathbf{y}}$ and that distribution, that is, $\frac{1}{n_y}\sum_{i=1}^{n_y} \mathcal{D}^2(\mathbf{y}_i, \mathcal{X}) = \mathcal{D}^2(\overline{\mathbf{y}}, \mathcal{X})$. Parameters $\{\overline{\mathbf{x}}, \boldsymbol{\Sigma}_x\}$ belonging to $\mathcal{X}$ were estimated using the sample $\{\mathbf{x}_j\}_{j=1,\ldots,n_x}$.

[1] Usually 3 in a color coordinates such as *RGB*, *HSI* or *Lab*.

Therefore, the reverse distance $\mathcal{D}^2\left(\overline{\mathbf{x}},\mathcal{Y}\right)$ between $\{\mathbf{x}_i\}$ points and the distribution $\mathcal{Y}$, whose parameters are computed using the sample $\{\mathbf{y}_i\}$, can also be taken into account. Hence, it seems natural to define the total distance between the two distributions $\mathcal{X}$ and $\mathcal{Y}$ as the mean of the two previous distances, that is, $\mathcal{D}^2(\mathcal{X},\mathcal{Y}) = \frac{1}{2}(\mathcal{D}^2(\overline{\mathbf{x}},\mathcal{Y}) + \mathcal{D}^2(\overline{\mathbf{y}},\mathcal{X}))$, which is equivalent to the expression

$$\mathcal{D}^2\left(\mathcal{X},\mathcal{Y}\right) = \left(\overline{\mathbf{x}}-\overline{\mathbf{y}}\right)\left[\frac{1}{2}\left(\varSigma_x^{-1}+\varSigma_y^{-1}\right)\right]\left(\overline{\mathbf{x}}-\overline{\mathbf{y}}\right)^t \tag{3}$$

Fréchet Distance. Another way to compute a distance between two MGDs is the Fréchet distance [13]. Fréchet distance between two random variables X and Y is defined by $\min_{X,Y} E\{\| X-Y \|^2\}$. This is a special case of the Monge–Kantorovich mass transference problem. Dowson and Landau [13] solved this problem for the case of $\mathcal{X}$ and $\mathcal{Y}$ being *elliptically symmetric*, which is the condition of the MGD. Hence, the distance between $\mathcal{X}$ and $\mathcal{Y}$ can be written as

$$\mathcal{D}^2\left(\mathcal{X},\mathcal{Y}\right) = \|\overline{\mathbf{x}}-\overline{\mathbf{y}}\|^2 + tr\left[\mathbf{\Sigma}_x+\mathbf{\Sigma}_y-2\left(\mathbf{\Sigma}_x\mathbf{\Sigma}_y\right)^{1/2}\right] \tag{4}$$

Fréchet distance is composed of two terms, namely, an Euclidean distance among means and a distance on the space of covariance matrices. Additionally, it is a closed–form solution to the Earth's Mover Distance (EMD) in the situation of two equally weighted Gaussian[2] and a natural distance for the Gaussian region representation.

Fröbenius Distance. Alike Fréchet distance, this one computes the distance between two MGD by addition of two partial distances, one among means (Euclidean) and another among covariances, which is defined between matrices based on the norm of the difference matrix computed from the covariances $\mathbf{\Sigma}_x$ and $\mathbf{\Sigma}_y$, calculated as if they were vectors (componentwise), i.e.,

$$\mathcal{D}^2(\mathcal{X},\mathcal{Y}) = \|\overline{\mathbf{x}}-\overline{\mathbf{y}}\|^2 + \|\mathbf{\Sigma}_x-\mathbf{\Sigma}_y\|^2 \tag{5}$$

Bhattacharyya Distance. Bhattacharyya's affinity kernel [14] was extensively used as a similarity measure in tasks such as object tracking [15]. It is defined as $K(\mathcal{X},\mathcal{Y}) = \int_{\varOmega}\sqrt{P_x(\mathbf{v})P_y(\mathbf{v})}\,d\mathbf{v}$, where P_x and P_y are PDFs of the random variables X and Y, respectively, and $\mathbf{v}\in\varOmega\subset\mathbb{R}^d$. This is a divergence–type measure interpretable as a (normalized) correlation between PDFs [15].

A closed form in the case of MGDs for the above kernel is suggested in [16] as $K(\mathcal{X},\mathcal{Y}) = k\cdot exp\left(-\frac{1}{2}\mathcal{D}^2(\mathcal{X},\mathcal{Y})\right)$, where $k = |\mathbf{\Sigma}_z|^{1/2}/(|\mathbf{\Sigma}_x|^{1/4}|\mathbf{\Sigma}_y|^{1/4})$ and, by analogy, $\mathcal{D}^2\left(\mathcal{X},\mathcal{Y}\right)$ is

$$\mathcal{D}^2(\mathcal{X},\mathcal{Y}) = \frac{1}{2}\left(\overline{\mathbf{x}}\,\mathbf{\Sigma}_x^{-1}\,\overline{\mathbf{x}}^t + \overline{\mathbf{y}}\,\mathbf{\Sigma}_y^{-1}\,\overline{\mathbf{y}}^t - 2\,\overline{\mathbf{z}}\,\mathbf{\Sigma}_z^{-1}\,\overline{\mathbf{z}}^t\right) \tag{6}$$

with the additional definitions of matrix $\mathbf{\Sigma}_z = [\frac{1}{2}(\mathbf{\Sigma}_x^{-1}+\mathbf{\Sigma}_y^{-1})]$ and vector $\overline{\mathbf{z}} = \frac{1}{2}(\mathbf{\Sigma}_x^{-1}\overline{\mathbf{x}}+\mathbf{\Sigma}_y^{-1}\overline{\mathbf{y}})$.

[2] This assumption can be totally assumed whenever segmented images are employed.

Kullback–Leibler Distance. The Kullback–Leibler (KL) divergence is a measure of the alikeness between two PDFs based on information theoretic motivations [17], defined as $KL(\mathcal{X},\mathcal{Y}) = \int_{\Omega} P_x(\mathbf{v})\, log(P_x(\mathbf{v})/P_y(\mathbf{v}))\, d\mathbf{v}$. If both P_x and P_y are MGDs, it turns into $KL(\mathcal{X},\mathcal{Y}) = tr\left(\mathbf{\Sigma}_y^{-1}\mathbf{\Sigma}_x\right) + log|\mathbf{\Sigma}_y| - log|\mathbf{\Sigma}_x| - d$.

Since KL divergence is not symmetric in general, it must be symmetrized before defining a proper distance as follows $KLS(\mathcal{X},\mathcal{Y}) = \frac{1}{2}(KL(\mathcal{X},\mathcal{Y}) + KL(\mathcal{Y},\mathcal{X}))$. Consequently, the KLS distance thus obtained is given as

$$KLS(\mathcal{X},\mathcal{Y}) = \frac{1}{2}\left(tr\left(\mathbf{\Sigma}_y^{-1}\mathbf{\Sigma}_x\right) + tr\left(\mathbf{\Sigma}_x^{-1}\mathbf{\Sigma}_y\right) - 2d\right) \tag{7}$$

As for the cases of Fröbenius and Fréchet, Eq. (7) only represents a metric in the covariance space, so the distance is $\mathcal{D}^2(\mathcal{X},\mathcal{Y}) = \|\overline{\mathbf{x}} - \overline{\mathbf{y}}\|^2 + KLS(\mathcal{X},\mathcal{Y})$.

Jensen–Shannon Distance. KL divergence has a number of numerical difficulties when covariances are close to singularity. A variant to overcome such a problem is the *Jensen–Shannon divergence* (JSD), defined as $JSD(\mathcal{X},\mathcal{Y}) = \frac{1}{2}(KL(\mathcal{X},\frac{\mathcal{X}+\mathcal{Y}}{2}) + KL(\mathcal{Y},\frac{\mathcal{X}+\mathcal{Y}}{2}))$, which in the MGD case changes into

$$JSD(\mathcal{X},\mathcal{Y}) = \frac{1}{2}\left(tr\left(2\left(\mathbf{\Sigma}_x + \mathbf{\Sigma}_y\right)^{-1}\mathbf{\Sigma}_x\right) + tr\left(2\left(\mathbf{\Sigma}_x + \mathbf{\Sigma}_y\right)^{-1}\mathbf{\Sigma}_y\right)\right) \tag{8}$$

Again, Eq. (8) is just a distance between covariances and must be completed to get a distance between distributions as $\mathcal{D}^2(\mathcal{X},\mathcal{Y}) = \|\overline{\mathbf{x}} - \overline{\mathbf{y}}\|^2 + JSD(\mathcal{X},\mathcal{Y})$.

3.3 IRM Similarity Measure

Integrated Region Matching (IRM) measures the overall similarity between images by integrating distances among regions of two images. An advantage of the overall similarity measure is its robustness against poor segmentations. Precisely, a region–to–region match is obtained when regions are significantly similar to each other in terms of the extracted signature $\mathcal{X}_i$, i.e., the most similar regions are matched first. Then, the whole distance is computed as a weighted sum of distances between region pairs as follows

$$IRM(\mathcal{X},\mathcal{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{m} s_{ij}d_{ij} \tag{9}$$

where $d_{ij} = \mathcal{D}(\mathcal{X}_i,\mathcal{X}_j)$ and $s_{ij} \geq 0$ is the level of significance between two regions, which indicates the relevance of the matching for determining the whole distance between the two images. It is required that the most similar regions get the highest priority, so the IRM algorithm in [11] attempts to assign as much significance as possible to region pairs with the least distance d_{ij}.

Additionally, the selection of the region weights w_i must be faced. These values are related to the levels of significance as $\sum_j s_{ij} = w_i$ and $\sum_i s_{ij} = w'_j$. Despite that choice can be done *uniformly* for all regions, we prefer the *area percentage scheme*, where w_i is the ratio between region area and image area, since more salient objects in an image tend to occupy larger areas, besides of being less sensitive to inaccurate segmentations.

4 Experiments and Results

The database used in these experiments belongs to the *Columbia Object Image Library* (COIL)[3], which consists in the color images of 100 objects viewed under 72 poses against a black background. Nevertheless, we used a smaller set of $N_{set} = 18$ views per object to get greater variations, along with only $N_{obj} = 51$ objects – Fig. 1(a) –, which definitively makes a sufficient database of $N_{total} = 918$ color images to test the above distances between segmentations. For each image, two segmentations were obtained by using Figueiredo's EM in [2] and our graph–partition approach in [1]. In Fig. 1(b) and Fig. 1(c) we show some segmentation results of the objects' frontal views in Fig. 1(a).

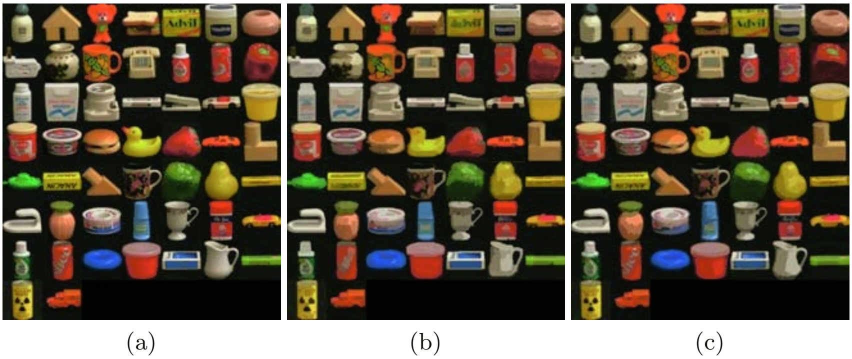
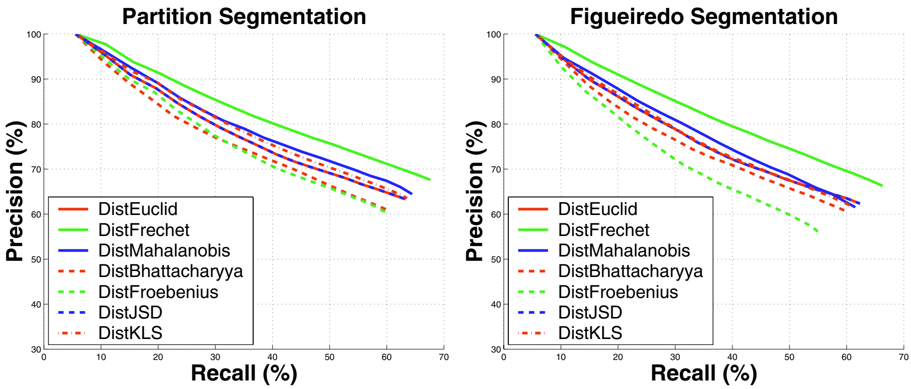


Fig. 1. COIL database: (a) original frontal views, (b) graph–partition segmentation, and (c) Figueiredo's segmentation.

We want to establish the performance of each of the aforementioned distances to benchmark them for any posterior application. To that purpose, we carry out two kinds of tests, an image retrieval and an object matching experiments.

Image retrieval emulates the response of a system to a global query. That response is a set of images sorted by their increasing distance to the queried image. After sorting them, only the first N_{ret} images are selected. $N_{ret} = 1, \ldots, N_{set}$ is the num. of images retrieved. The num. of relevant images retrieved is N_{rel}. Two measures are used to evaluate the retrieval performance, namely, recall and precision [12]. *Recall* is the percentage of the total relevant images retrieved, N_{rel}/N_{set}, whereas *precision* refers to the capability of the system to retrieve only relevant images, N_{rel}/N_{ret}. The total num. of relevant images are the num. of images per object set N_{set}. The *Precision vs. Recall* plots in Fig. 2 for each segmentation approach comprehensively exhibit those results. Those graphics show how precision decays when the fraction of relevant images is pushed up. The slower the fall, the better. Hence, Fréchet distance seems the best distance, while Fröbenius is the worst, for both segmentation algorithms. Besides, graph–partition results are slightly better than those of Figueiredo's.

[3] `http://www1.cs.columbia.edu/CAVE/research/softlib/coil-100.html`



(a) (b)

Fig. 2. Precision vs. Recall plots corresponding to each segmentation algorithm: (a) graph–partition and (b) Figueiredo's EM.

Matching experiment consists in evaluating every distance as a way to perform object identification. The former database was divided into two disjoint subsets containing 9 images per object, that is, 459 images per set in total. One set is the *testing* set while the other is the *sample* set. Then, the experiment computes the total amount of correct identifications carried out by taking images from the testing set and finding the closest image in the sample set. A correct matching occurs whenever the closest image recovered from the sample set belongs to the same object as the image from the testing set.

Table 1. Object recognition results per segmentation method.

Distance	**Figueiredo** [%]	**Partition** [%]
Euclidean	89.11 (4)	91.29 (3)
Fréchet	**94.34 (1)**	**95.21 (1)**
Mahalanobis	89.54 (2)	91.50 (2)
Bhattacharyya	86.71 (5)	88.02 (6)
Fröbenius	84.53 (6)	89.32 (5)
Jensen–Shannon	89.11 (4)	91.29 (3)
KL Symmetric	89.32 (3)	91.07 (4)

Results corresponding to object identification are exhibited in Table 1 performed using Lab color coordinates after the two previously mentioned color image segmentation algorithms, namely, graph–partition and Figueiredo's EM. In regard to the recognition rates, the best overall results were obtained using our segmentation and the Fréchet distance, and was as high as 95.21% of correct object identifications, outperforming Mahalanobis distance. Euclidean distance obtains medium positions, similarly to JSD and KLS, whereas the worst ones are both Fröbenius and Bhattacharyya.

5 Conclusions

This paper illustrated the problem of comparing images by means of their color segmentations. A group of seven distances were proposed within the frame of the IRM distance and the employ of Multivariate Gaussian Distributions (MGD) for the color description of image regions. The performance of these distances was examined in tasks such as image retrieval and object recognition using the two segmentation algorithms in [1] and [2]. The best overall results were obtained for both tasks using the graph–partition approach along with the Fréchet distance, outperforming other distances in comparing MGDs.

References

1. Vergés-Llahí, J., Climent, J., Sanfeliu, A.: Colour image segmentation solving hard-constraints on graph-partitioning greedy algorithm. In: Proc. 15^{th} International Conference on Pattern Recognition , ICPR'00. Volume 3. (2000) 629–632
2. Figueiredo, M., Jain, A.: Unsupervised learning of finite mixture models. IEEE Trans. on Pattern Analysis and Machine Intelligence **24** (2002) 381–396
3. Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Dim, B., Huang, Q., Gorkani, M., Hafner, J., Lee, D., Petkovick, D., Steele, D., Yanker, P.: Query by image and video content: The QBIC system. Computer **28** (1995) 23–32
4. Pentland, A., Picard, R., Sclaroff, S.: Photobook: Tools for content-based manipulation of image databases. Proc. SPIE **2185** (1994) 34–47
5. Gupta, A., Jain, R.: Visual information retrieval. Comm. ACM **40** (1997) 69–79
6. Smith, J., Chang, S.: VisualSEEK: A fully automated content-based query system. In: Proc. ACM Multimedia. (1996) 87–98
7. Wang, J., Wiederhold, G., Firschein, O., Sha, X.: Content-based image indexing and searching using Daubechies' wavelets. Int'l Digital Libraries **1** (1998) 311–328
8. Natsev, A., Rastogi, R., Shim, K.: WALRUS: A similarity retrieval algorithm for images databases. IEEE Trans. on Knowledge and Data Eng. **16** (2004) 301–316
9. Ma, W., Manjunath, B.: NeTra: A toolbox for navigating large image database. In: Proc. IEEE Int'l Conf. Image Processing. (1997) 568–571
10. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: Image segmentation using expectation-maximization and its application to image querying. IEEE Trans. on Pattern Analysis and Machine Intelligence **24** (2002) 1026–1038
11. Wang, J., Li, J., Wiederhold, G.: SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. IEEE Trans. on Pattern Analysis and Machine Intelligence **23** (2001) 947–963
12. Nascimento, M., Sridhar, V., Li, X.: Effective and efficient region-based image retrieval. Journal of Visual Languages and Computing **14** (2003) 151–179
13. Dowson, D., Landau, B.: The Fréchet distance between multivariate normal distributions. Journal of Multivariate Analysis **12** (1982) 450–455
14. Bhattacharyya, A.: On a measure of diverg. between two statistical populations defined by their probability distrib. Bull. Calcutta Math. Soc. **35** (1943) 99–110
15. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Trans. on Pattern Analysis and Machine Intelligence **25** (2003) 564–577
16. Kondor, R., Jebara, T.: A kernel between sets of vectors. In: Proc. Int. Conf. on Machine Learning, ICML 2003. (2003)
17. Kullback, S.: Information Theory and Statistics. Dover, New York (1968)

An Algorithm for the Detection of Multiple Concentric Circles

Margarida Silveira

IST/INESC, Lisbon, Portugal

Abstract. This paper presents a method for the detection of multiple concentric circles which is based on the Hough Transform (HT). In order to reduce time and memory space the concentric circle detection with the HT is separated in two stages, one for the center detection and another for the radius determination. A new HT algorithm is proposed for the center detection stage which is simple, fast and robust. The proposed method selects groups of three points in each of the concentric circles to solve the circle equation and vote for the center. Geometrical constraints are imposed of the sets of three points to guarantee that they in fact belong to different concentric circles. In the radius detection stage the concentric circles are validated. The proposed algorithm was compared with several other HT circle detection techniques. Experimental results show the superiority and effectiveness of the proposed technique.

1 Introduction

There are several applications that require the automatic detection of concentric circles. Some examples include iris detection, the detection of washers in industrial parts or the detection of inhibition halos of antibacterial activity in the food industry.

However, and despite the fact that circle detection with the Hough Transform (HT) was been widely studied, the specific case of the detection of concentric circles has received little attention. In fact, the only publication on this matter known to the authors is the one proposed by Cao and Deravi [9]. The Cao-Deravi method is based on scanning the image horizontally and vertically to search for groups of three edge points in order to determine the parameters of a circle. The edge gradient direction is used as a guide, together with a set of search rules, to select the three edge points. The fact that this method only performs horizontal and vertical scanning makes it inadequate when there is occlusion or deformation of the circles. In addition, and even though gradient direction is not used directly, the fact that is used in the search makes the method susceptible to noise because gradient direction is very much affected by noise.

The method we propose for concentric circle detection is also based on the search for three edge points belonging to the same circle. However, we scan the image in various directions and instead of using gradient direction to guide the search we rely on geometrical constraints. In the HT there is a tradeoff between computational effort in the edge space and computational effort in the accumulator space. The method we

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 271–278, 2005.

propose involves more work in the edge space and is therefore more effective in situations where the background is complex or the image noise is significant, because in those cases the analysis of the accumulator is too complex.

The proposed method will be described in section 2.

2 Detection of Concentric Circles

Circles are described by three parameters, the center coordinates (*a,b*) and the radius *r*:

$$(x-a)^2+(y-b)^2=r^2 \tag{1}$$

Therefore, the conventional HT [1] needs a three dimensional parameter space. In order to reduce time and memory space, the problem of circle detection was separated in two stages [2] [3]. The first stage involves a two parameter HT to find the center (*a,b*) of the circles and the second stage involves a one dimensional HT, that is, a simple histogram, to identify the radius of the outer circle. A new algorithm is proposed for the center detection stage. Since there are several circles in the image, there will be several peaks in the center parameter space corresponding to the different circle centers. In addition, there may be spurious peaks resulting from the interference of different circles. Therefore, the one dimensional Hough transform designed to identify the radius will also be used to validate the existence of two, or more, concentric circles.

We assume that the position and number of concentric circles halos is variable and unforeseeable and that the size of the concentric circles lies within the a known range $r_{\min}$ to $r_{\max}$.

2.1 Detection of Circle Centers

We propose a HT center detection method that is simple and, as results will show, is fast and reliable. The algorithm is robust to noise since it does not use gradient information and is able to detect irregular and partially occluded circles.

After edge detection, the connected components are labeled. For each point $A=(x_A,y_A)$ another two points $B=(x_B,y_B)$ and $C=(x_C,y_C)$ of the same component are randomly selected that satisfy the following expressions:

$$d_{\min}^2 \le (x_A-x_B)^2+(y_A-y_B)^2 \le 4{r_{\max}}^2 \tag{2}$$

$$d_{\min}^2 \le (x_A-x_C)^2+(y_A-y_C)^2 \le 4{r_{\max}}^2 \tag{3}$$

$$d_{\min}^2 \le (x_B-x_C)^2+(y_B-y_C)^2 \le 4{r_{\max}}^2 \tag{4}$$

where $r_{\max}$ is the maximum value allowed for the radius and $d_{\min}$ prevent selecting points too close.

The three points A, B and C are used to solve the circle equation and find a candidate circle center. Let $O = (x_O, y_O)$ denote that candidate center.

In the case of two concentric circles, the lines between AO, BO and CO should intersect a different connected component at points D, E and F, respectively or even at points G, H and I. Moreover the angle between $\overline{BA}$ and $\overline{BC}$ should be the same as the angle between $\overline{ED}$ and $\overline{EF}$ or $\overline{HI}$ and $\overline{HG}$. This is illustrated in Fig. 1.

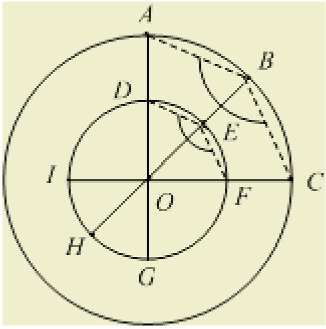


Fig. 1. The lines between each point A, B and C on the outer circle and the center should intersect the inner circle at points D, E and F respectively. The angle between line segments $\overline{BA}$ and $\overline{BC}$ equals the angle between $\overline{ED}$ and $\overline{EF}$.

Therefore, the proposed algorithm will draw three lines that go through each of the points A, B, C and the candidate center O and conduct a search for edge points that lie on each of the three lines and belong to a given connected component which is different from the one that points A, B and C belong to. If such three points D, E and F are found and they verify the angle requirement, $\phi\left(\overline{BA},\overline{BC}\right) = \phi\left(\overline{ED},\overline{EF}\right)$, then the center coordinates O are incremented in the two dimensional Hough space. Additionally, the three new points are used to find another estimate of the center O and this new estimate is also incremented in the Hough accumulator in order to increase the center detection accuracy.

The Bresenham line drawing algorithm [4] was used to draw the lines and the Cramer's rule was used to solve the circle equation from the sets of three points. The angle between two line segments, for instance $\overline{BA}$ and $\overline{BC}$ is calculated by the following expression:

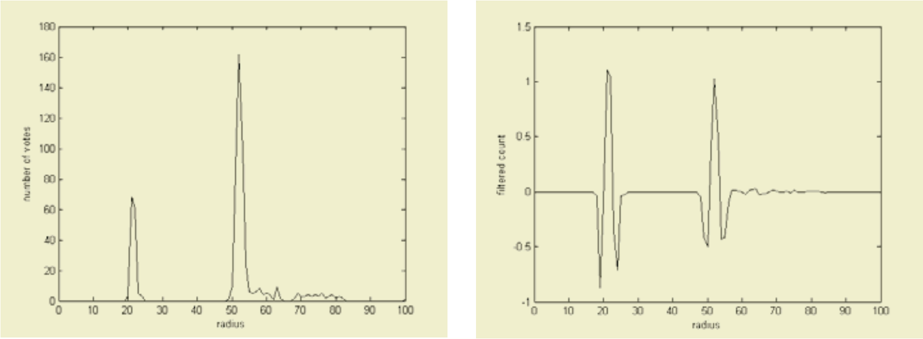
$$\phi\left(\overline{BA},\overline{BC}\right) = arc\cos\frac{\overline{BA}\ .\ \overline{BC}}{\left\|BA\right\|\left\|BC\right\|} \tag{5}$$

The number of concentric circles is variable and unforeseeable, and consequently the centers accumulator will have several peaks. In the radius detection stage described in section 2.2 each candidate center will be validated. After a circle has been analyzed, in the centers accumulator a small region of cells around that center are zeroed and then the accumulator is searched for the next peak.

2.2 Detection of Circle Radius

After a candidate circle center has been detected, edge points in the $2r_{\max}$ square region around the candidate center vote for the corresponding circle radius using expression (1).

Concentric circles will originate several peaks in the radius accumulator corresponding to the different radii. Since larger concentric circles will also originate higher peaks in the radius accumulator, the count is normalized. In addition, the accumulator is filtered to enable the detection of the more diffuse or deformed circles. The filter proposed in [9] was used. We require that circles have some percentage of their circumference appearing in the image. That percentage may be different for the inner circles than for the outer ones depending on the application.



a) b)

Fig. 2. Radius histogram. Two maxima appear corresponding to two concentric circles. The identified circles were at $r = 22$ and $r = 52$ a) Number of votes b) Filtered count.

A group of two concentric circles is identified as two peaks in the radius accumulator that verify the percentage requirement. Fig. 2 shows an example of the radius histogram corresponding to an accepted concentric circle. In case there is prior knowledge about the expected ring width, that information can be incorporated to check the separation between peaks. Naturally, the method can be extended to deal with any number of concentric circles.

3 Experimental Results

In this section we will present results of the proposed method using real and synthetic images. We will study the performance of the technique with increasing additive Gaussian noise and compare it with the Ioannou method proposed in [9] and the Cao-Deravi method [9]. The Cao-Deravi method was described in the introduction. The Iaonnou method, although it was not specifically designed for the detection of concentric circles, it can be utilized for that purpose if the radius validation suggested in section 2.2 is used. This method exploits the property that every line that perpendicularly bisects any chord of a circle passes through its centre. Therefore, it selects pairs

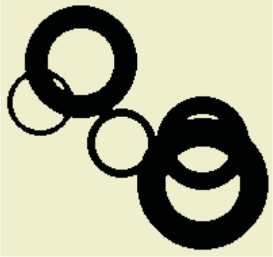


Fig. 3. Test image example.

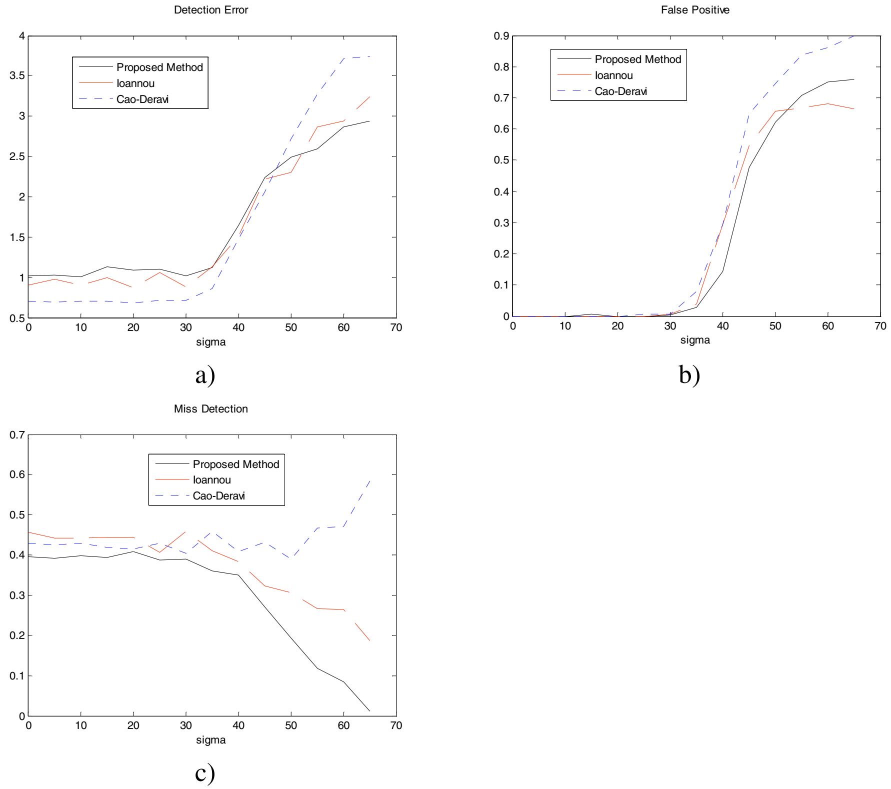
of points of the same connected component and finds the line that perpendicularly bisects the two points. All the points on this line that belong to the parameter space are incremented. This method is very accurate and robust when compared to methods that rely on edge gradient direction. The Bresenham line drawing algorithm [4] was used to find the bisection line.

The tests were performed in the following conditions. For each value of sigma 50 test images of size 256x256 were created. Each image consisted of 5 black rings on a white background. The position, size and width of the rings were all randomly selected meaning there may be overlap. An example is show in Fig. 3. The Canny operator was used for edge detection because of its good localization and its robustness in the presence of noise, and also because Canny´s gradient magnitude is not as sensitive to contour orientation as other detectors [7]. In the Cao-Deravi algorithm edge direction was obtained with Wilson and Bhalerao's operator [5]. The radius threshold was set to 0.3 for both the outer and the inner circle. All the angles calculated with expression (**5**) used PI/60 accuracy.

The different techniques were compared as to their false positive rate, miss detection rate and accuracy. The radius threshold value has an influence on these statistics. In fact, increasing this threshold would increase the miss detection rate and simultaneously decrease the false positive rate.

The results are presented in Fig. 4. As it would be expected, the performance of all the methods decreases with increasing noise sigma. It can be seen that the method has much lower miss detection rate than the other methods and also less false positive rate, although with slightly worse accuracy when there is little noise. The Cao-Deravi method has the best accuracy for low amounts of noise but is also the less resistant to increasing noise.

Fig. 5 shows an example of the application of the different methods on a real image with a circular traffic sign and a complicated background. The images show that the proposed method was the best in detecting the traffic sign. The Cao-Deravi method was also able to detect the traffic sign although slightly skewed. The Ioannou method failed to detect the sign and produced two false concentric circles. Another example is shown in Fig. 6 where the test image is a motorcycle. In this example both the proposed and the Ioannou methods detected the front tire and the results are quite similar. The Cao-Deravi method however, missed the front tire and detected the fender and also originated a false detection.



a)

b)

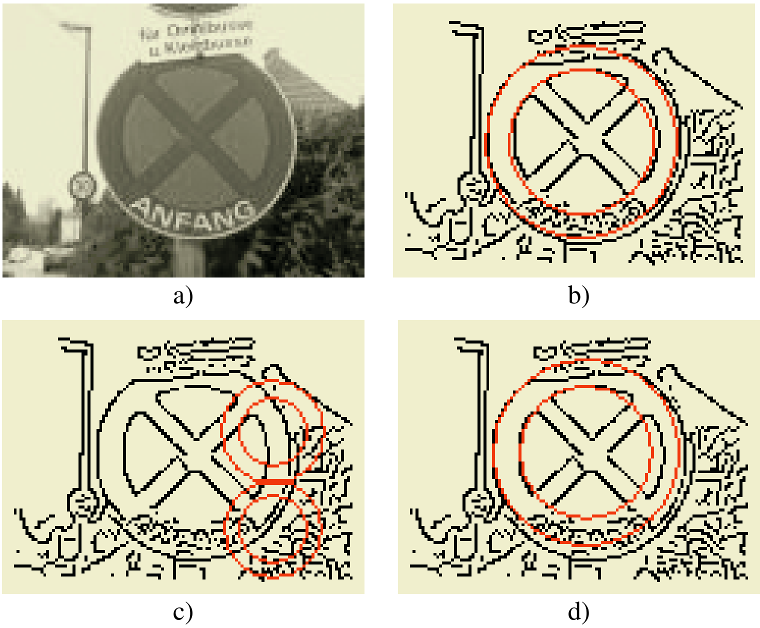
c)

Fig. 4. Comparison of the methods performance with increasing noise. a) Detection error b) False positive rate c) Miss detection rate.

4 Conclusions

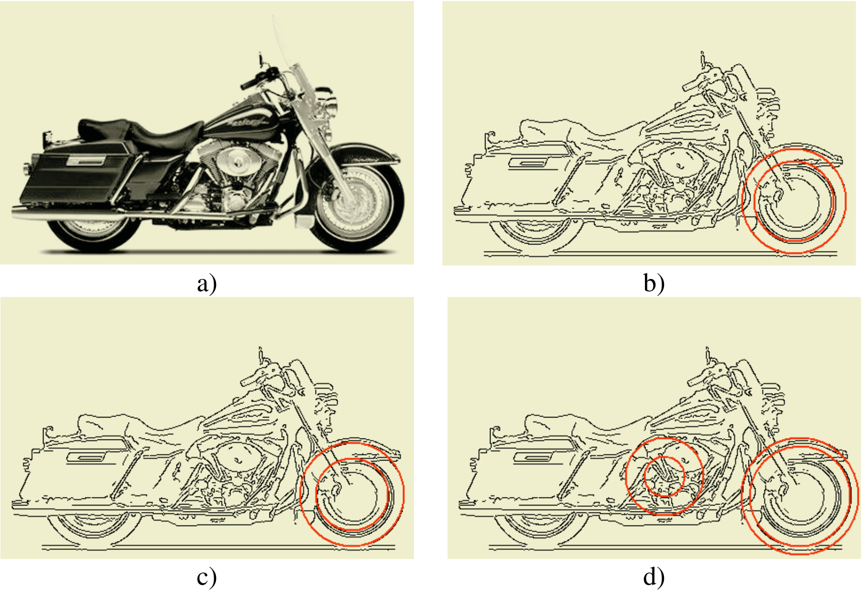
This paper proposed a method for the detection of multiple concentric circles. In order to reduce time and memory space the circle detection with the HT was separated in two stages, one for the center detection and another for the radius determination. A new algorithm was proposed for the detection of the circle centers. The proposed method selects groups of three points in each of the concentric circles to solve the circle equation and vote for the center. Geometrical constraints are imposed of the sets of three points to guarantee that they in fact belong to different concentric circles. The search is performed in several directions and doesn't rely on the use of edge direction. Therefore the algorithm is robust to noise and occlusion. The radius determination stage also performs a validation of the concentric circles.

Examples were provided to illustrate the performance of the algorithm. The proposed algorithm was favorably compared with other HT center detection methods. Experiments showed that the method is more robust to noise.



a) b) c) d)

Fig. 5. Example of the methods performance using a real image. a) Original image b) Results of the proposed method superimposed on the edge map c) Results of the Ioannou method superimposed on the edge map d) Results of the Cao-Deravi method superimposed on the edge map.



a) b) c) d)

Fig. 6. Example of the methods performance using a real image. a) Original image b) Results of the proposed method superimposed on the edge map c) Results of the Ioannou method superimposed on the edge map d) Results of the Cao-Deravi method superimposed on the edge map.

References

1. Duda, R., Hart, P., Use of the Hough Transformation to Detect Lines and curves in Pictures. Communications of the ACM, 15, 1, (1972), 11-15
2. Illingworth, J., Kittler, J.: The Adaptive Hough Transform. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-9, 5 (1987) 690–698
3. Kimme, C., Ballard, D.H., Sklansky, J.: Finding Circles by an Array of Accumulators. Communications of the ACM, 2, (1975), 120-122
4. Foley D., J., van Dam, A., Feiner, S., K., Hughes, J.,F.: Computer Graphics: Principles and Practice. 2nd edn., Addison-Wesley, Reading, MA (1997)
5. Wilson, R., Bhalerao, H.: Kernel Designs for Efficient Multiresolution Edge Detection and Orientation Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 14, 3 (1992) 384-389
6. Haralick, R., M., Shapiro, L.G.: Computer and Robot Vision. Addison-Wesley (1992) 80-81
7. Ziou, D.: The Influence of Edge Direction on the Estimation of Edge Contrast and Orientation. Pattern Recognition, 34, 4 (2001) 855–863
8. Ioannou, D., Huda, W., Laine, A.F.: Circle Recognition through a 2D Hough Transform and Radius Histogramming. Image and Vision Computing, 17, 1 (1999) 15–26
9. Cao, X., Deravi, F.: An Efficient Method for the Detection of Multiple Concentric Circles. Proc. International Conference on Acoustic, Speech and Signal Processing, ICASSP'92, San Francisco, (1992) III-137-III-140

Image Corner Detection Using Hough Transform

Sung Kwan Kang, Young Chul Choung, and Jong An Park

Dept. of Information & Communications Engineering,
Chosun University, Gwangju, Korea
japark@chosun.ac.kr

Abstract. This paper describes a new corner detection algorithm based on the Hough Transform. The basic idea is to find the straight lines in the images and then search for their intersections, which are the corner points of the objects in the images. The Hough Transform is used for detecting the straight lines and the inverse Hough Transform is used for locating the intersection points among the straight lines, and hence determine the corner points. The algorithm was tested on various test images, and the results are compared with well-known algorithms.

Keywords: corner detection, Hough Transform, Detecting Straight Lines, curvature scale, corner points.

1 Introduction

Corners have been found to be very important in human perception of shapes and have been used extensively for shape description, recognition, and data compression [1]. Corner detection is an important aspect of image processing and finds many practical applications. Applications include motion tracking, object recognition, and stereo matching. Corner detection should satisfy a number of important criteria. It should detect all the true corners, and the corner points should be well localized, and should be robust with respect to noise, and should be efficient. Further, it should not detect false corners.

There is an abundance of literature on corner detection. Moravec [2] observed that the difference between the adjacent pixels of an edge or a uniform part of the image is small, but at the corner, the difference is significantly high in all directions. Harris [3] implemented a technique referred to as the Plessey algorithm. The technique was an improvement of the Moravec algorithm. Beaudet [4] proposed a determinant (DET) operator which has significant values only near corners. Dreschler and Nagel [5] used Beaudet's concepts in their detector. Kitchen and Rosenfeld [6] presented a few corner-detection methods. The work included methods based on gradient magnitude of gradient direction, change of direction along edge, angle between most similar neighbors, and turning of the fitted surface. Lai and Wu [7] considered edge-corner detection for defective images. Tsai [8] proposed a method for boundary-based corner detection using neural networks. Ji and Haralick [9] presented a technique for corner detection with covariance propagation. Lee and Bien [10] applied fuzzy logic to corner detection. Fang and Huang [11] proposed a method which was an improvement on the gradient magnitude of the gradient-angle method by Kitchen and Rosenfeld. Chen and Rockett utilized Bayesian labeling of corners using a gray-level corner image model in [12]. Wu and Rosenfeld [13] proposed a technique which examines

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 279–286, 2005.

the slope discontinuities of the x and y projections of an image to find the possible corner candidates. Paler et al. [14] proposed a technique based on features extracted from the local distribution of gray-level values. Rangarajan et al. [15] proposed a detector which tries to find an analytical expression for an optimal function whose convolution with the windows of an image has significant values at corner points. Arrebola et al. [16] introduced corner detection by local histograms of contour chain code. Shilat et al. [17] worked on ridge's corner detection and correspondence. Nassif et al. [18] considered corner location measurement. Sohn et al. [19] proposed a mean field-annealing approach to corner detection. Zhang and Zhao [20] considered a parallel algorithm for detecting dominant points on multiple digital curves. Kohlmann [21] applied the 2D Hilbert transform to corner detection. Mehrotra et al. [22] proposed two algorithms for edge and corner detection. The first is based on the first-directional derivative of the Gaussian, and the second is based on the second-directional derivative of the Gaussian. Davies [23] applied the generalized Hough transform to corner detection. Zuniga and Haralick [24] utilized the facet model for corner detection. Smith and Brady [25] used a circular mask for corner detection. No derivatives were used. Orange and Groen [26] proposed a model-based corner detector. Other corner detectors have been proposed in [27-30]. Mokhtarian [31] used the curvature-scale-space (CSS) [32], [33] technique to search the corner points. The CSS technique is adopted by MPEG-7. The Kitchen and Rosenfeld detector [6], the SUSAN detector [25] and the CSS [32] corner detector have shown good performance. These detectors are therefore chosen as our test detectors.

In this paper, a new corner detection based on the forward and inverse Hough Transform is presented. The straight lines in the images are detected and their intersection points are used to locate the corner points. This paper is organized as follows. Section 2 describes the method to determine the straight lines in the images using Hough Transform and section 3 describes the algorithm to detect the intersection points among the straight lines using the inverse Hough Transform. Section 4 describes the simulation results. At the end, we will conclude our paper with few final remarks.

2 Detecting Corners Using Inverse Hough Transform

Ideally, a corner is an intersection of two straight lines. However, in practice, corners in the real world are frequently deformed with ambiguous shapes. As corner represent certain local graphic features at abstract level, corners can intuitively be described by some semantic patterns (see Fig. 1). A corner can be characterized as one of the following four types:

- Type A: A perfect corner as modeled in [30], i.e., a sharp turn of curve with smooth parts on both sides.
- Type B: The first of two connected corners similar to the END or STAIR models in [30], i.e., a mark of change from a smooth part to a curved part.
- Type C: The second of two connected corners, i.e., a mark of change from a curved part to a smooth part.
- Type D: A deformed model of type A, such as a round corner or a corner with arms neither long nor smooth. The final interpretation of the point may depend on the high level global interpretation of the shape.

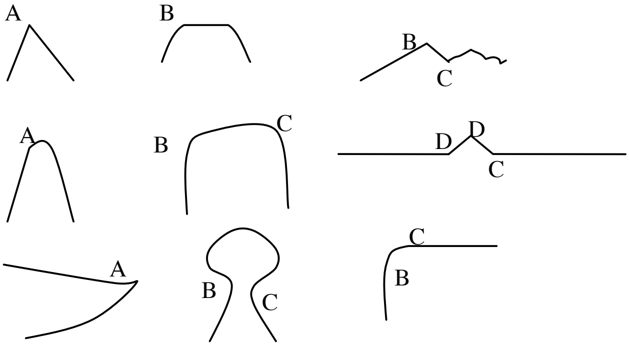


Fig. 1. Four types of corners.

Figure 2 shows some examples of the four types of the corner. It is obvious from the Fig.2 that the corner points at very small level are the intersection points of the two straight lines. To detect the intersection points between the straight lines, and hence the corner points in an image, we do the following procedure.

- Quantise (x,y) space into a two-dimensional array C for the original size of the image in appropriate steps of x and y.
- Initialise all elements of C(x,y) to zero.
- For each pixel (ρ’,θ’) in parameter space, we add 1 to all elements of C(x,y) whose indices x and y satisfy $\rho' = x\cos\theta' + y\sin\theta'$.
- Search for elements of C(x,y) which have large values than one. Each one found corresponds to a possible candidate for corners in the original image.

Because of many intersections of lines, false corners are also detected. To avoid false candidates, the detected corners whose vicinity does not contain any edge point are discarded. Now consider the case of a corner in an image as shown in Fig. 2.

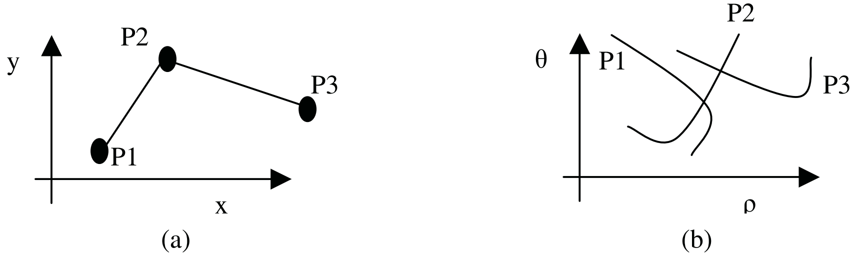


Fig. 2. (a) Points in image space (b) Corresponding points in parameter space.

The corner formed by three points P1, P2 and P3 are transformed into (ρ, θ) parameter space. In parameter space, corner point P2 has two intersections with other lines, while P1 and P3 have only one. It means P2 is a corner point. The peaks in the parameter space correspond to the corners in the image space. We get many intersection points in the image space after getting the inverse transform of the peaks from the parameter space. To remove the unwanted intersection points (i.e., no corner points), we ANDED the intersection points with the edges of the image. The position of true corners and the edges will coincide and give the actual position of the corners. To get more accurate results and to avoid large number of intersections, corners are detected block by block processing with sliding overlapping window. The number of computations becomes higher, but the results are more accurate.

3 Simulations

The proposed algorithm was tested using three different images, the same images used in [31]. The results are compared with three different corner detectors, namely, Kitchen and Rosenfeld, SUSAN and CSS corner detectors. The results for the three corner detectors are also taken from [31]. The results for each detector were the best results obtained by searching the best parameters. The three test images show in Fig .3, Fig.4, and Fig.5 are called Blocks, House and Lab. The Blocks test image contains much texture and noise. The House image has a lot of small details and texture in the brick wall. The Lab image contains plenty of corners. The results show that the proposed algorithm gives better result than that of KR and SUSAN method and gives comparable results with that of CSS.

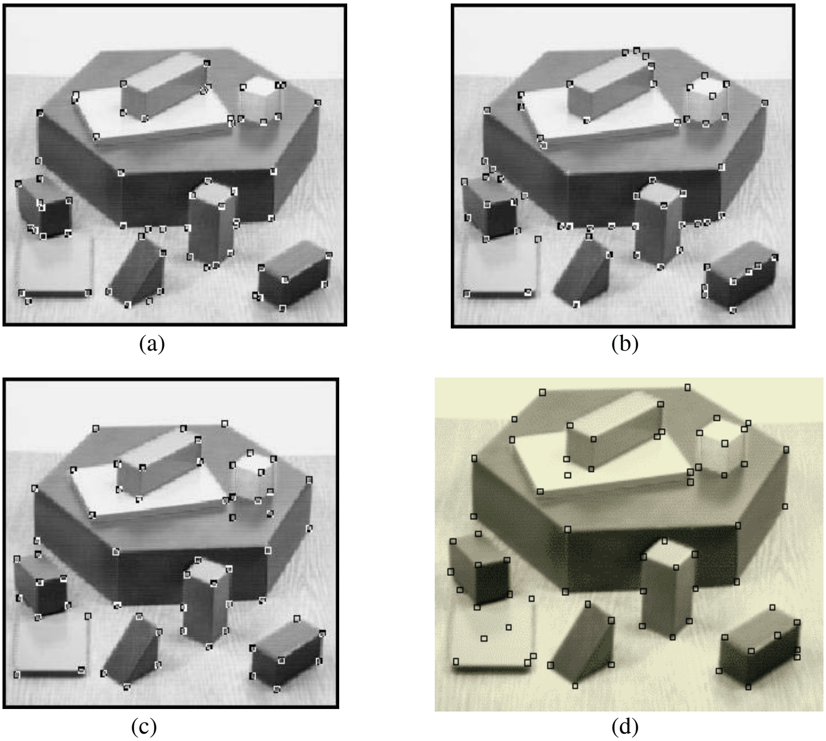
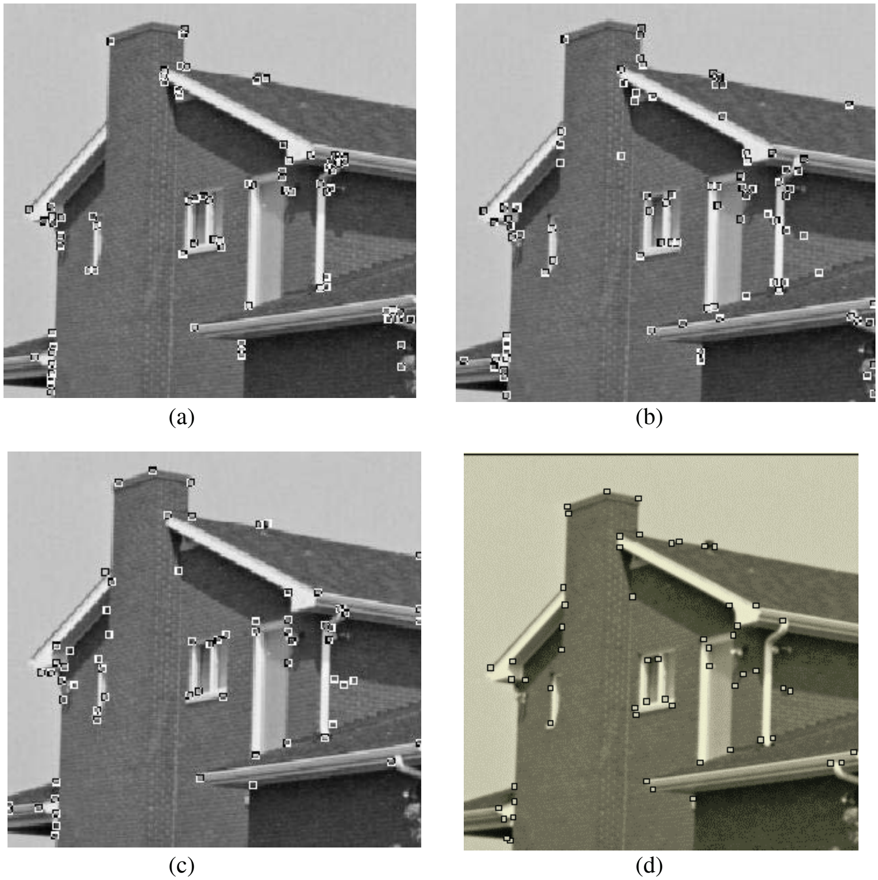


Fig. 3. Blocks image. (a) Kitchen/Rosenfeld. (b) SUSAN. (c) CSS. (d) Proposed Algorithm.



(a)

(b)

(c)

(d)

Fig. 4. House image. (a) Kitchen/Rosenfeld. (b) SUSAN. (c) CSS. (d) Proposed Algorithm.

The proposed method and CSS perform well on the Blocks image. Other detectors find difficulties in locating the corner points. Similarly, the proposed method and CSS method show good performance on the House and the Lab images, while others perform badly. Overall our proposed method and CSS are comparable. The most of the time of the proposed method is consumed by the Hough Transforms. By decreasing or increasing the criterion for peaks effect the detail information about the corners.

4 Conclusions

In this paper, a new corner detector based on the Hough Transform is proposed. The edges in the image are found and the edges are transformed from the image space to parameter space. The Hough Transform is used to find the straight lines in the image, and the inverse Hough Transform is used to find the intersection points among the straight lines. The intersection points are the corner points. The proposed method is compared with the previous methods. The results are comparable with the curvature scale space image corner detector.

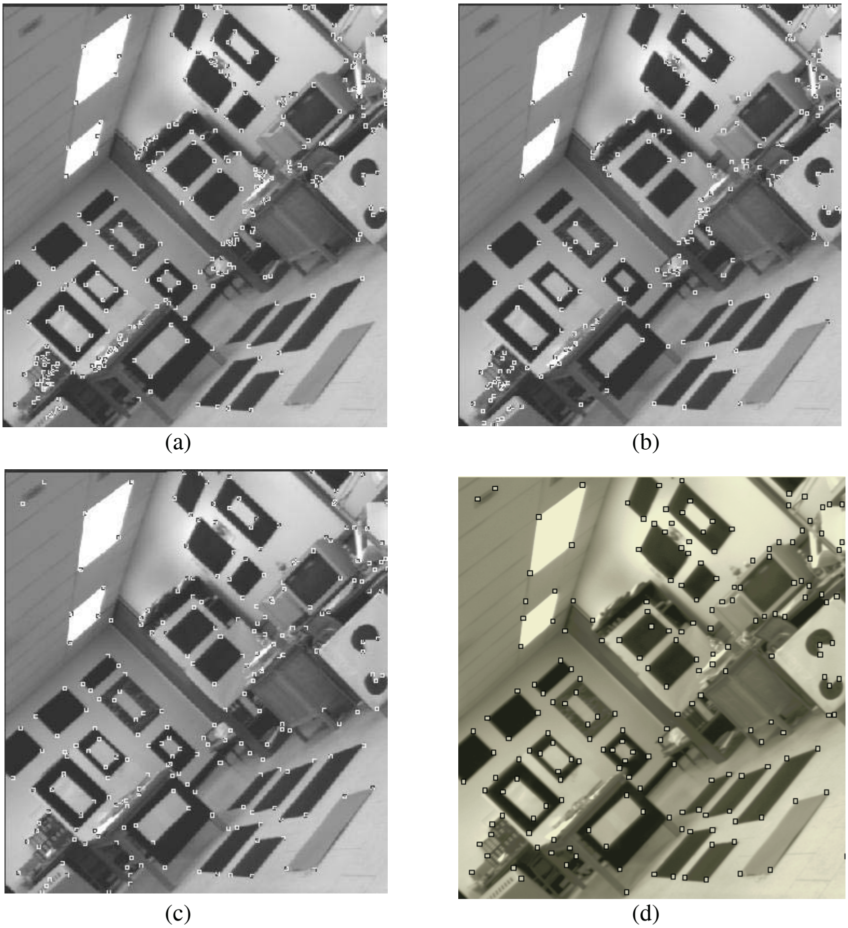


Fig. 5. Lab image. (a) Kitchen/Rosenfeld. (b) SUSAN. (c) CSS. (d) Proposed Algorithm.

Acknowledgement

This research was supported by the Program for the Training of Graduate Students in Regional Innovation which was conducted by the Ministry of Commerce, Industry and Energy of the Korean Government

References

1. Liyuan Li and Weinan Chen, “Corner detection and interpolation on planar curves using fuzzy reasoning,” IEEE Trans. On Pattern Analysis and Machine Vision, vol. 21, no. 11, November, 1999.
2. H. P. Moravec, “Towards automatic visual obstacle avoidance,” Proc. Int’l Joint Conf. Artificial Intelligence, p. 584, 1977.

3. C. G. Harris, “Determination of ego-motion from matched points,” Proc. Alvey Vision Conf., Cambridge, UK, 1987.
4. P. R. Beaudet, “Rotationally invariant image operators,” Int’l Joint Conf. Pattern Recognition, pp. 579-583, 1978.
5. L. Dreschler and H. H. Nagel, “Volumetric model and 3D trajectory of a moving car derived from monocular TV frame sequences of a street scene,” Int’l Joint Conf. Artificial Intelligence, pp. 692-697, 1981.
6. L. Kitchen and A. Rosenfeld, “Gray level corner detection,” Pattern Recognition Letters, pp. 95-102, 1982.
7. K. K. Lai and P. S. Y. Wu, “Effective edge-corner detection method for defected images,” Proc. Int’l Conf. Signal Processing, vol. 2, pp. 1,151-1,154, 1996.
8. D. M. Tsai, “Boundary based corner detection using neural networks,” Pattern Recognition, vol. 30, no. 1, pp. 85-97, 1997.
9. Q. Ji and R. M. Haralick, “Corner detection with covariance propagation,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 362-367, 1997.
10. K. J. Lee and Z. Bien, “Grey-level corner detector using fuzzy logic,” Pattern Recognition Letters, vol. 17, no. 9, pp. 939-950, 1996.
11. J. Q. Fang and T. S. Huang, “A corner finding algorithm for image analysis and registration,” Proc. AAAI Conf., pp. 46-49, 1982.
12. W. C. Chen and P. Rockett, “Bayesian labeling of corners using a grey-level corner image model,” IEEE Int’l Conf. Image Processing, vol. 1, pp. 687-690, 1997.
13. Z. O. Wu and A. Rosenfeld, “Filtered projections as an aid to corner detection,” Pattern Recognition, vol. 16, no. 31, 1983.
14. K. Paler, J. Foglein, J. Illingworth, and J. Kittler, “Local ordered grey levels as an aid to corner detection,” Pattern Recognition, vol. 17, no. 5, pp. 535-543, 1984.
15. K. Rangarajan, M. Shah, and D. Van Brackle, “Optimal corner detector,” Computer Vision, Graphics, and Image Processing, vol. 48, pp. 230-245, 1989.
16. F. Arrebola, A. Bandera, P. Camacho, and F. Sandoval, “Corner detection by local histograms of contour chain code,” Electronics Letters, vol. 33, no. 21, pp. 1,769-1,771, 1997.
17. E. Shilat, M. Werman, and Y. Gdalyahu, “Ridge’s corner detection and correspondence,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 976-981, 1997.
18. S. Nassif, D. Capson, and A. Vaz, “Robust real-time corner location measurement,” Proc. IEEE Conf. Instrumentation and Measurement Technology, pp. 106-111, 1997.
19. K. Sohn, J. H. Kim, and W. E. Alexander, “Mean field annealing approach to robust corner detection,” IEEE Trans. Systems, Man, and Cybernetics, vol. 28B, no. 1, pp. 82-90, 1998.
20. X. Zhang and D. Zhao, “Parallel algorithm for detecting dominant points on multiple digital curves,” Pattern Recognition, vol. 30, no. 2, pp. 239-244, 1997.
21. K. Kohlmann, “Corner detection in natural images based on the 2-D Hilbert Transform,” Signal Processing, vol. 48, no. 3, pp. 225-234, 1996.
22. R. Mehrotra, S. Nichani, and N. Ranganathan, “Corner detection,” Pattern Recognition, vol. 23, no. 11, pp. 1,223-1,233, 1990.
23. E. R. Davies, “Application of the generalized Hough Transform to corner detection,” IEE Proc., vol. 135, pp. 49-54, 1988.
24. O. A. Zuniga and R. M. Haralick, “Corner detection using the facet model,” Proc. Conf. Pattern Recognition and Image Processing, pp. 30-37, 1983.
25. S. M. Smith and J. M. Brady, “SUSAN—A new approach to low level image processing,” Defense Research Agency, Technical Report no. TR95SMS1, Farnborough, England, 1994.
26. C. M. Orange and F. C. A. Groen, “Model based corner detection,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, 1993.

27. Bejamin Bell and L. F. Pau, "Contour tracking and corner detection in a logic programming environment," IEEE Trans. On Pattern Analysis and Machine Vision, vol. 12, no. 9, September 1990.
28. H. C. Liu and M. D. Srinath, "Corner detection from chain-code," Pattern Recognition, vol. 23, nos. ½, pp. 51-68, 1990.
29. H. Freeman and L. Davis, "A corner-finding algorithm for chain-coded curves," IEEE Trans. Coputers, vol. 26, pp. 297-303, 1977.
30. A. Rattarrangsi and R. T. Chin, "Scale-based detection of corners of planar curves," IEEE Trans. On Pattern Analysis and Machine Vision, vol. 14, no. 4, pp. 430-449, 1992.
31. Farzin Mokhtarian and Riku Suomela, "Robust image corner detection through curvature scale space," IEEE Trans. On Pattern Analysis and Machine Vision, vol. 20, no. 12, December, 1998.
32. F. Mokhtarian and A.K. Mackworth, "A theory of Multi-Scale, Curvature-based shape representation for planar curves," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 14, no. 8, pp. 789-805, Aug. 1992.
33. F. Mokhtarian and R. Suomela, "Curvature Scale Space for robust image corner detection," Int'l Conf. Pattern Recognition, Brisbane, Australia, 1998.

Dissimilarity Measures for Visual Pattern Partitioning*

Raquel Dosil[1], Xosé R. Fdez-Vidal[2], and Xosé M. Pardo[1]

[1] Dep. de Electrónica e Computación, Univ. de Santiago de Compostela,
Campus Universitario Sur, s/n, 15782, Santiago de Compostela, Spain
rdosil@usc.es, pardo@dec.usc.es
http://www-gva.dec.usc.es/grupo/grupo.htm
[2] Escola Politécnica Superior, Univ. de Santiago de Compostela,
Campus Universitario, s/n, 27002, Lugo, Spain
faxose@usc.es
http://www.lugo.usc.es/~ffeeps

Abstract. We define a *visual pattern* as an image feature with frequency components in a range of bands that are aligned in phase. A technique to partition an image into its visual patterns involves clustering of the band-pass filtered versions of the image according to a measure of congruence in phase or, equivalently, alignment in the filter's responses energy maxima. In this paper we study some measures of dissimilarity between images and discuss their suitability to the specific task of misalignment estimation between energy maps.

1 Introduction

The identification and extraction of relevant low level features in an image is of great importance in image analysis. Field states that meaningful features present some degree of alignment in the phase of its spectral components [1]. In the RGFF representational model introduced in [2] and extended in [3, 4], such features are called *visual patterns* and defined as patterns with alignment in a set of local statistics along wide frequency ranges. These methods can detect a wide variety of features, like textures, grating patters, blobs and symmetric and antisymmetric discontinuities in intensity, texture, and phase. They share a common scheme consisting of the decomposition of the image into elementary features using a bank of log Gabor filters followed by the clustering of these features according to some measure of dissimilarity among them.

The distance used in [2] and [3] is inspired in biological processes. It combines attention mechanisms and pooling of sensor outputs. Attention points are identified as energy maxima. On their part, Dosil et al. [4] use a distance based on the normalized mutual information of the filter's response energy, which is less computationally expensive, less parameterized and less dependent on the performance of low level processes like non-maxima suppression and scale estimation. Mutual information I is

* The authors desire to acknowledge the Xunta de Galicia for their financial support of this work by means of the research project PGIDIT04TIC206005PR.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 287–294, 2005.

widely employed as a measure of image dissimilarity in various fields of application, with great popularity in medical image registration [5, 6]. However, we have observed that the behavior of *I* is not completely satisfactory for filter clustering. In some cases it groups very dissimilar frequency features. This is due to that *I* treats intensity values qualitatively, increasing with the concurrence of weak and strong maxima. *I* is an underconstrained measure of dependency since it makes no assumptions about the kind of functional relation between the images – see [7] for a detailed explanation.

Then again, a measure that allows a generic dependency between the images may not be the most appropriate in all applications. In the specific case of filter's responses energy maps it seems that the kind of dependency that best reflects the relation between features belonging to the same visual pattern is linear functional. A measure of similarity that constrains the allowed relations between two images to a linear transformation is the correlation coefficient. To test this assumption, here we make a comparison among a series of dissimilarity measures, including distances based on correlation coefficient, mutual information and the original measure proposed in the RGFF.

In the next section the set of dissimilarity measures between pairs of filtered images is presented. In section 3, the method for visual pattern partitioning is described. Section 4 presents an experimental study on the performance of these measures in the task of visual pattern partitioning. Section 5 presents the conclusions derived from it.

2 Dissimilarity Between Energy Maps

All measures presented here are derived from a similarity measure δ by applying to it a transformation to enhance intercluster distances, invert its range and map it to the interval [0, 1]. What follows is the list of proximities δ and their correspondent distances D_{δ}. *X* and *Y* represent energy maps and *M* is number of bins in an histogram.

a) Normalized mutual information [4, 5]
If *H* stands for entropy, then

$$NI(X,Y) = 2 \cdot \frac{I(X,Y)}{H(X)+H(Y)}, \quad \text{where } I(X,Y) = H(X)+H(Y)-H(X,Y).$$

$$D_{NI}(X,Y) = \left(1-\sqrt{NI(X,Y)}\right)^2. \tag{1}$$

b) Correlation ratio η [7, 8, 9]

$$\eta^2(X \mid Y) = 1 - \mathrm{Var}(X - \mathrm{E}(X \mid Y))/\mathrm{Var}(X)$$

$$D_{\eta}(X,Y) = 1-\sqrt{\max\left(\eta^2(X \mid Y), \eta^2(Y \mid X)\right)} \tag{2}$$

c) Correlation coefficient
This measure has into account the sign of the correlation coefficient, so that an image and its inverse have maximum distance

$$\rho(X,Y) = \mathrm{Cov}(X,Y) \Big/ \sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$$
$$D_\rho(X,Y) = \left(1 - \sqrt{(1+\rho(X,Y))/2}\right)^2 \tag{3}$$

e) Toussaint's distance [9, 10]

$$T(X,Y) = \sum_{i,j} P_{x,y}(i,j) - \frac{2P_{x,y}(i,j)P_x(i)P_y(j)}{P_{x,y}(i,j) + P_x(i)P_y(j)}$$
$$D_T(X,Y) = \left(1 - \sqrt{T(X,Y)/T_{\max}}\right)^2, \quad \text{with} \quad T_{\max} = 1 - 2/(M+1) \tag{4}$$

f) Lin's K divergence [9, 10]

$$K_{div}(X,Y) = \sum_{i,j} P_{x,y}(i,j) \log \frac{2P_{x,y}(i,j)}{P_{x,y}(i,j) + P_x(i)P_y(j)}$$

$$D_{Kdiv}(X,Y) = \left(1 - \sqrt{K_{div}(X,Y)/K_{div\ \max}}\right)^2, \text{ with } K_{div\ \max} = \log\left(2M/(M+1)\right) \tag{5}$$

g) Dissimilarity measures on energy maxima.
A new dissimilarity measure D^*_δ is obtained from each D_δ as follows

$$D^*_\delta(X,Y) = D_\delta(X',Y') \tag{6}$$

where *X'* and *Y'* are respectively *X* and *Y* after non-maxima suppression. Maxima are determined by comparing each point with its neighbors in the filter's direction.

h) RGFF dissimilarity measure [2]
For each energy map *X*, the set of its maxima Ω_X is determined. For each *p* in Ω_X a vector T^p of length *Q* of local statistics is measured. Then, for a given $\beta > 0$

$$D_\beta(X,Y) = \frac{1}{Card(\Omega_X)}\Big(\sum_{p\in\Omega_X} \left|\mu_p(X,Y)\right|^\beta \Big)^{1/\beta}, \mu_p(X,Y) = \sum_{k=1}^{Q} \frac{1}{\omega_k} d\left(T_k^p(X), T_k^p(Y)\right).$$
$$D_{RGFF}(X,Y) = D_\beta^2(X,Y) + D_\beta^2(Y,X). \tag{7}$$

where ω_X is the maximum T_k over all Ω_X and all *X*. The local statistics they employ are local phase, normalized local energy and its entropy, contrast and standard deviation.

2.1 Computational Cost of Dissimilarity Estimation

One of the main advantages of global measures in relation to the RGFF measure is their lower computational cost. In the following, an analysis of the asymptotic computational cost of the presented approaches is presented.

Let us suppose that the input data are a volume of dimensions $N \times N \times N$, that our filter bank consists of *F* filters and that the number of bins used for histogram calculations is *M*. The calculus of ρ is $O(N^3)$ while the estimation of *NI*, η, *T* and K_{div}

involves the construction of the joint histogram of the two maps, which is $O(N^3)$, and the posterior accumulation of the contributions of each bin in the histogram, which is $O(M^2)$. Supposing that N and M are of the same order of magnitude, the cost of the dissimilarity calculation is $O(N^3)$. This must be done for each of the $F(F-1)$ pairs of filters, resulting in a computational cost of $O(F^2 \cdot N^3)$.

In the case of the RGFF distance, the cost of the dissimilarity calculations is $O(F^2 \cdot N^6)$. This is due to the calculus of the neighborhood of each attention point and the local statistics on it. The neighborhoods are related to the scales of each maximum and are defined as the distance from each energy maxima to the nearest minimum. In high scale filters the neighborhood radius is in the order of the image size. Hence, this calculations are $O(N^3)$ and must be done for each attention point, i.e., $O(N^3)$ times, and for each filter pair, i.e., $O(F^2)$ times. Even if the points of each neighborhood where stored, what would have a memory cost of $O(F \cdot N^6)$, the calculus of the local statistics differences maintains a total cost $O(F^2 \cdot N^6)$.

3 Visual Pattern Partitioning Methodology

Visual pattern partitioning of a 3D image consists of the next sequence of steps:

1. Selection of *active* – with high information content
2. Calculation of the energy maps correspondent to the *active* filters' responses
3. Measure of dissimilarity between pairs of energy maps
4. Hierarchical clustering of the energy maps based on the dissimilarity matrix
5. Visual pattern reconstruction by linear summation of cluster energy maps.

In the next subsections these procedures are detailed.

3D Filter Bank

The filters' transfer function T is designed as the product of separable factors R and S in the radial and angular components respectively with expressions

$$R(\rho;\rho_i) = \exp\left\{-\log^2(\rho/\rho_i)\big/\left(2\log^2(\sigma_\rho/\rho_i)\right)\right\}, \tag{8}$$

where σ_ρ is the standard deviation and ρ_i the central radial frequency and

$$S(\phi,\theta;\phi_i,\theta_i) = S(\alpha) = \exp\left\{-\alpha^2/(2\sigma_\alpha^2)\right\}, \text{ with } \alpha(\phi_i,\theta_i) = \mathrm{acos}\left(\mathbf{f}\cdot\mathbf{v}/\|\mathbf{f}\|\right), \tag{9}$$

where $\mathbf{v} = (\cos\phi_i \cos\theta_i, \cos\phi_i \sin\theta_i, \sin\phi_i)$ is a unit vector in the filter's direction, σ_α is the angular standard deviation and $\mathbf{f}$ the point in the frequency space in Cartesians.

In our configuration elevation is sampled uniformly, while azimuth is non-uniformly sampled by maintaining equal arc-length between adjacent azimuth values over the unit radius sphere. The bank has been designed using 4 elevations –only one hemisphere is needed due to symmetry– and 6 azimuths to sample half the $z = 0$ plane, yielding 23 orientations with angular bandwidth of 25°. In the radial axis, 4 values have been taken with wavelengths 4, 8, 16 and 32 and 2 octave bandwidth.

Selection of Active Bands

To decrease the computational cost, the number of filters is reduced by discarding filters with wavelengths greater than half the image size, roughly representing the average intensity, and with low information content, named *non active*. The measure of information density is $E = \log (|F| + 1)$, where F is the image Fourier transform.

A band is *active* if it comprises any value of E over the maximum spectral noise. The maximum noise level is estimated as $m + x\sigma$, where m is the mean noise energy, σ is its standard deviation and $x \geq 0$. Here, m and σ have been measured in the band of frequencies greater that double the largest of the bank's central frequencies and $x = 3$.

To eliminate remaining spurious noise "spots" a radial median operator is applied, which only considers neighbors that are anterior or posterior in the radial direction to calculate the median. This eliminates isolated peaks but preserving the continuity of structures along scales. In this work the mask size is taken to $L = 3$.

Feature Clustering

Here, hierarchical clustering has been chosen to group features, using a complete-link algorithm, where the distance between clusters is defined as the maximum of all pairwise intercluster distances, thus producing compact groups. The number of clusters N_c is an input parameter of the algorithm. The usual strategy to determine the optimal N_c is to run the algorithm for each possible N_c and evaluate the quality of each resulting partition according to a given validity index. Here, the modified Davies-Boulding index, introduced in [9] has proved to produce good results. It is a graph-theory based index that measures the compactness of the clusters in relation to their separation.

4 Results

To compare the performance of the presented dissimilarity measures, the visual pattern partitioning method described in section 3 has been applied using each of them to a set of test images. The test bench is composed of 32 images, 13 of them 2D and the other 19 3D. While it is quite easy to determine if the results obtained for a 2D image are correct by visual inspection, this is more difficult for 3D images. For this reason, all the 3D images in the bench are synthetic. The correctness of the results is determined by comparing them with the design specifications. The result must contain one cluster for each visual pattern in the image and their frequency bands must match the expected ones. 2D cases are either synthetic images or natural images with clearly identifiable visual patterns or images synthesized as a collage of natural Brodatz textures. In this last type the result must contain one cluster for each texture. Additionally, they may appear patterns correspondent to texture boundaries.

The results obtained are summarized in Fig. 1. The measures have been sorted by percentage of correct results. It can be seen that D_{ρ} has the best performance, followed by D_{NI}. In general the results are not very good due to the complexity of the

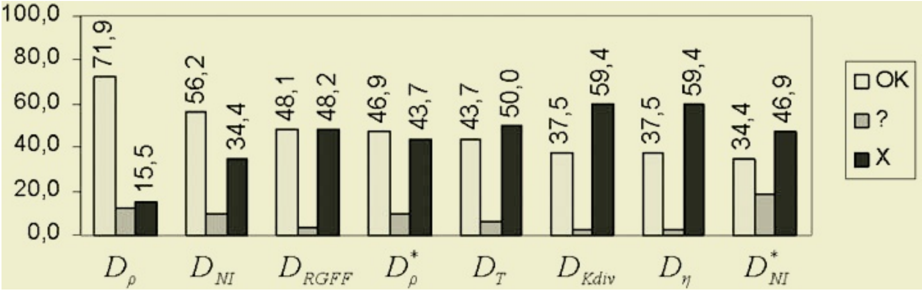


Fig. 1. Percentage of correct (OK), incorrect (X) and indecisive (?) results for each distance

task of matching same-pattern frequency features, which present strong differences – these results can not be extrapolated to other applications, like image registration.

Fig. 2 shows one example result for a 3D image that presents diverse visual patterns: a grating pattern, a plane – even feature – and a phase change – odd feature. In this specific case all the distances produce the correct result, except from D_T and D^*_{NI} which do not detect the phase change.

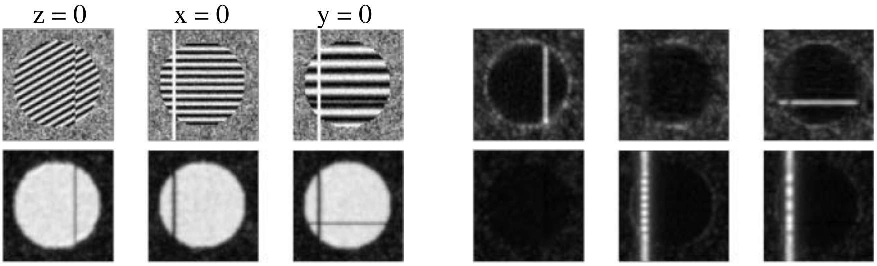


Fig. 2. *Top left*: Cross sections of the original 3D data. *Remainder*: Sections of the three patterns isolated using D_ρ

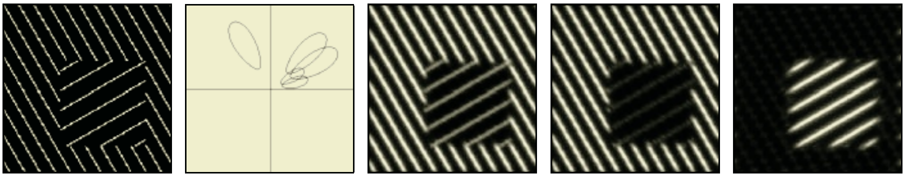


Fig. 3. *From left to right:* Original image. One of the clusters obtained with D^*_{NI}, represented by the $e^{-1/2}$ level curves of the filters' transfer function. Pattern associated to the previous cluster. The two patterns obtained with D_ρ

The remainder figures illustrate the improvements brought by the use of the D_ρ distance in relation to the other measures. Fig. 3 shows an example of 2D synthetic image. It can be seen that the D^*_{NI} distance groups orthogonal grating patterns together while D_ρ separates them. This is caused by the non null response of the filters to patterns with orientation orthogonal to it. Given that *NI* does not consider the magnitude of the difference between the responses of the filters, the resulting dissimilarity is small. Fig. 4 shows a similar example with a natural image of a Brodatz texture.

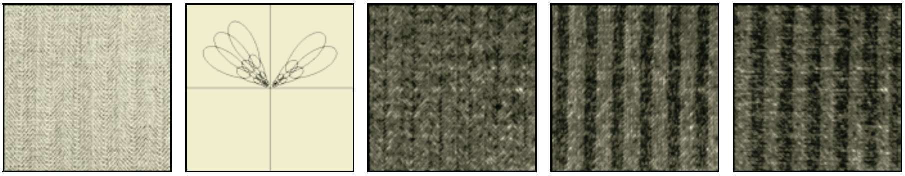


Fig. 4. *From left to right:* Original image. One of the clusters obtained with D_{NI}, represented by the $e^{-1/2}$ level curves of the filters' transfer function. Pattern associated to the previous cluster. The two patterns obtained with D_ρ

Fig. 5 and Fig. 6 present other cases were D_ρ corrects D_{NI} results. In Fig. 5 mutual information is not able to separate the texture inside the circle. Instead it decomposes the texture of the outer region into its vertical and horizontal components. In the example of Fig. 6 the results for D_{NI} are not shown since they are a total of 7 clusters, as the different components of each region have not been correctly integrated.

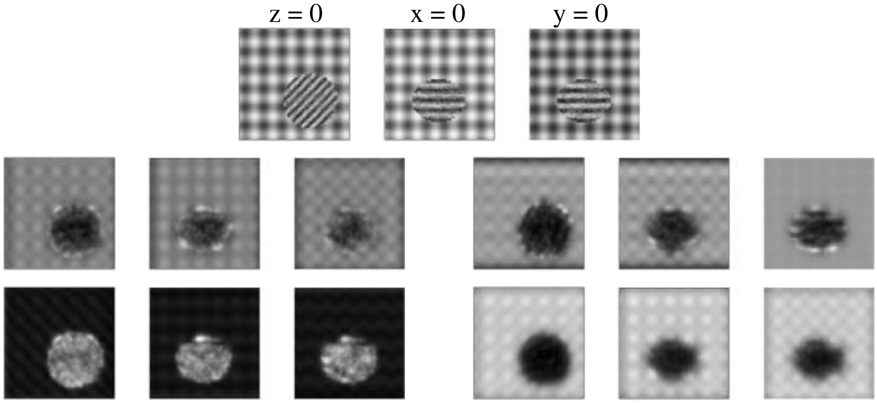


Fig. 5. *Top*: Cross sections of the original 3D data. *Middle*: Sections of the two patterns isolated using D_{NI}. *Bottom*: Sections of the two patterns isolated using D_ρ

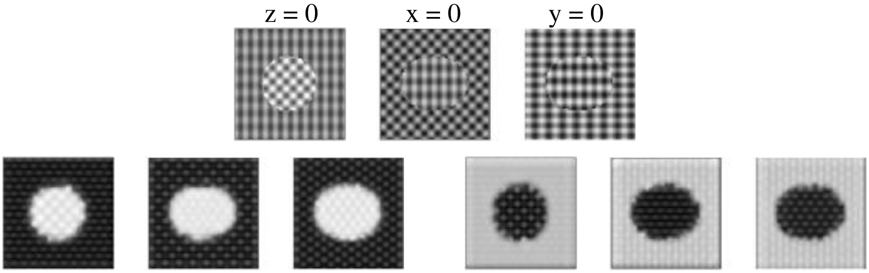


Fig. 6. *Top*: Cross sections of the original 3D data. *Bottom*: The two patterns isolated using D_ρ

5 Conclusions

Visual pattern partitioning makes reference to the process of isolation of the constituent low level features that are perceptually relevant in an image. It consists of the clustering of the frequency components of the image according to some distance reflecting the degree of alignment between them. In this paper we have discussed the suitability of a set of dissimilarity measures to this task.

We have planted the assumption that the kind of dependency that appears between the frequency components of the same visual pattern is a linear functional one. This explains the incorrect results obtained with measures based on mutual information, other information divergences and correlation ratio. Upon this assumption we predict that a measure based on the correlation coefficient should yield better results.

To test this hypothesis the dissimilarities have been tested with a set of 2D and 3D images. The results obtained have shown that the correlation coefficient distance solves the problems observed with mutual information and other global distances and improves the original measure proposed in the RGFF in speed and performance.

References

1. Field, D.J.: Scale–Invariance and self-similar “wavelet” Transforms: An Analysis of Natural Scenes and Mammalian Visual Systems. In: Farge, M., Hunt, J.C.R., Vassilicos, J.C. (eds.): Wavelets, fractals and Fourier Transforms, Clarendon Press, Oxford (1993) 151-193
2. Rodríguez-Sánchez, R., García, J.A., Fdez-Valdivia, J., Fdez-Vidal, X.R.: The RGFF Representational Model: A System for the Automatically Learned Partition of “Visual Patterns” in Digital Images, IEEE Trans. Pattern Anal. Mach. Intell., **21**(10) (1999) 1044-1073
3. Chamorro-Martínez, J., Fdez-Valdivia, J.A., García, J.A., Martínez-Baena, J.: A frequency Domain Approach for the Extraction of Motion Patterns, in IEEE International Conference on Acoustics, Speech and Signal Processing, Hong Kong, **3** (2003) 165-168
4. Dosil, R., Fdez-Vidal, X. R. and Pardo, X. M.: Multiresolution Approach to Visual Pattern Partitioning of 3D Images. In Campilho, A. and Kamel, M., (eds.) LNCS 3211: Image Analysis and Recognition, Porto, Portugal (2004) 655-663
5. Studholme, C. Hill, D. L. G. and Hawkes, D. J.: An Overlap Invariant Entropy Measure of 3D Medical Image Alignment. Pattern Recognition, **32** (1999) 71-86
6. Viola, P. A.: Alignment by Maximization of Mutual Information. Massachusetts Institute of Technology, Artificial Intelligence Laboratory, Technical Report 1548 (1995)
7. Roche, A., Malandain, G. and Ayache, N.: Unifying Maximum Likelihood Approaches in Medical Image Registration. Int. J. of Imaging Systems and Technology, **11** (2000) 71-80
8. Roche, A., Malandain, G., Pennec, X. and Ayache, N.: The Correlation Ratio as a New Similarity Measure for Multimodal Image Registration. In LNCS 1496: MICCAI’98, Springer-Verlag (1998) 115-1124
9. Sarrut, D. and Miguet S. Similarity Measures for Image Registration. 1st European Workshop on Content-Based Multimedia Indexing, Toulouse, France (1999) 263-270
10. Basseville, M.: Information: Entropies, divergences et moyennes. Technical Report 1020, IRISA, 35042 Rennes Cedex France (1996)
11. Pal, N.R., Biswas, J.: Cluster Validation Using graph Theoretic Concepts. Pattern Recognition, **30**(6) (1997) 847-857

A Comparative Study of Highlights Detection and Elimination by Color Morphology and Polar Color Models

Francisco Ortiz, Fernando Torres, and Pablo Gil

Automatics, Robotics and Computer Vision Group
Dept. Physics, Systems Engineering and Signal Theory, University of Alicante,
P.O. Box 99, 03080 Alicante, Spain
{fortiz,Fernando.torres,Pablo.Gil}@ua.es

Abstract. In this paper, we present a comparative study of detection and elimination of highlights in real color images of any type of material. We use different polar color spaces for the automatic highlight detection (HLS, HSV and *L1-norme*). To eliminate the highlights detected, we use a new connected vectorial filter of color mathematical morphology which it operates exclusively on bright zones, reducing the high cost of processing of the connected filter and avoiding over-simplification. The new method proposed here achieves good results and it not requires costly multiple-view systems or stereo images.

1 Introduction

In visual systems, images are acquired in work environments in which illumination plays an important role. Sometimes a bad adjustment of the illumination can introduce highlights (brightness or specular reflectance) into the objects captured by the vision system. Highlights in images have long been disruptive to computer-vision algorithms. They appear as surface features, which can lead to problems, such as stereo mismatching, false segmentation and recognition errors. In spite of such undesirable effects, however, there is, so far, no application available in commercial software that allows the automatic detection and elimination of such specularities.

To effectively eliminate the highlights in captured scenes, we must first identify them. The dichromatic reflection model, proposed by Safer [1], is one tool that has been used in many methods for detecting specularities. It supposes that the interaction between the light and any dielectric material produces different spectral distributions within the object (specular and diffuse reflectance). The specular reflectance has the same spectral makeup as the incident light, whereas, the diffused component is a product of illumination and surface pigments. Based on this model, Lin *et al* [2] have developed a system for eliminating specularities in image sequences by means of stereo correspondence. Bajcsy *et al* [3] use a chromatic space based on polar coordinates that allows the detection of specular and diffuse reflections by means of the previous knowledge of the captured scene. Klinker *et al* [4] employ a pixel-clustering

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 295–302, 2005.

algorithm that has been shown to work well in detecting brightness in images of plastic objects.

These above-mentioned approaches have produced good results but they entail requirements that limit their applicability, such as the use of stereo or multiple-view systems, long processing time, the previous knowledge of the scene, etc. Furthermore, some techniques merely detect brightness without eliminating it.

The organization of this paper is as follows: In Section 2, we present the color models for processing, together with the bi-dimensional diagrams used to detect the specular reflectance (highlights). In Section 3, we show the extension of the geodesic operations to color images. Section 4 we present the detection of highlights and our experimental results. The elimination process is presented in Section 5. Finally, our conclusions are outlined in the last section.

2 Polar Representations for Highlight Detection

In the last years, the color spaces based in polar coordinates are widely used in image processing. Important advantages of these color spaces are: good compatibily with human intuition of colors and separability of chromatic values from achromatic values. The two classic polar spaces are HLS and HSV [5,6]. The HLS and HSV components are calculated from RGB system. The formulas change from cartesian co-ordinates to polar co-ordinates. The luminance *l* and saturation *s* of HLS are calculated as follows:

$$\begin{cases} l = \dfrac{\max(r,g,b) + \min(r,g,b)}{2} \\ s = \begin{cases} \dfrac{\max(r,g,b) - \min(r,g,b)}{\max(r,g,b) + \min(r,g,b)} & if\ l \leq 0.5 \\ \dfrac{\max(r,g,b) - \min(r,g,b)}{2 - \max(r,g,b) + \min(r,g,b)} & if\ l > 0.5 \end{cases} \end{cases} \tag{1}$$

where *r*, *g*, *b* of RGB, and *s* and *l* range from 0 to 1. The HLS representation has cylindrical shape with the previous formulas. The HSV has a cone shape and the co-ordinates of value *v* and saturation *s* are calculated as follows:

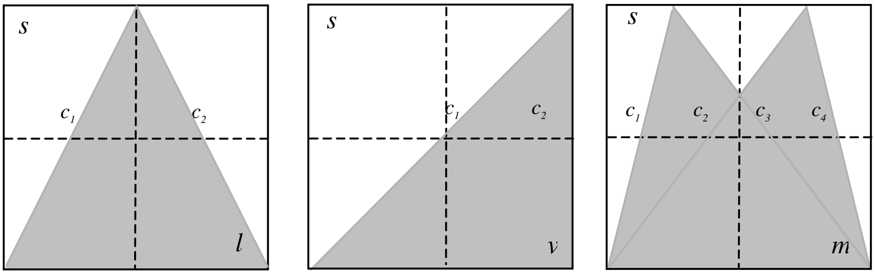
$$\begin{cases} v = \max(r,g,b) \\ s = \dfrac{\max(r,g,b) - \min(r,g,b)}{\max(r,g,b)} \end{cases} \tag{2}$$

We propose to exploit the existing relation of the specularities presents in a color image with specific coordinates of *l* and *s* in HLS or *v* and *s* in HSV, independently of the hue of the object in which the highlights appears [7]. These representations of colors are bi-dimensional histograms with a grey value, in which each co-ordinate (*l*,*s*) or (*v*,s) indicates the amount of pixels with *l* and *s* (HLS) or *v* and *s* (HSV) in the original color image. Figure 1.a and Figure 1.b show the LS diagram and VS diagram for HLS and HSV color spaces, respectively.

The two polar systems (HLS and HSV) have some in-coherences that prevent the use of these color representations in some image processing. This is due to the conversion formulas from RGB to polar spaces. Some instability arises in saturation of HLS and HSV spaces for small variations of RGB values. In addition, the saturation of the primary colors is not visually agreeable in HLS and HSV. In order to avoid these inconveniences we also use the Serra's *L1-norme* [8], where the intensity (achromatic) signal m and saturation signal s are calculated as follows:

$$\begin{cases} m = \dfrac{1}{3}(r+g+b) \\ s = \begin{cases} \dfrac{1}{2}(2r-g-b) = \dfrac{3}{2}(r-m) & \textit{if } (b+r) \geq 2g \\ \dfrac{1}{2}(r+g-2b) = \dfrac{3}{2}(m-b) & \textit{if } (b+r) < 2g \end{cases} \end{cases} \tag{3}$$

Figure 1.c shows the MS diagram from Serra's *L1-norme* as a positive projection of all the corners of the RGB cube in a normalization of the achromatic line to the m signal.



(a) (b) (c)

Fig. 1. 2D bi-dimensional diagrams of HLS (a), HSV (b) and *L1-norme* (c). Positive projection.

3 Vector Connected Filters

Morphological filters by reconstruction have the property of suppressing details while preserving the contours of the remaining objects [9,10]. The use of such filters in color images requires an ordered relationship among the pixels of the image. For the vectorial morphological processing, the lexicographical ordering o_{lex}= achromatic value → saturation → hue**,** will be used [11,12].

Once the orders have been defined, the morphological operators for the reconstruction of color images can be applied. Geodesic dilation is an elementary geodesic operation. Let $\boldsymbol{g}$ denote a marker color image and $\boldsymbol{f}$ a mask color image (if $o_{\mathrm{lex}}(\boldsymbol{g}) \leq o_{\mathrm{lex}}(\boldsymbol{f})$, then $\boldsymbol{g} \wedge_V \boldsymbol{f} = \boldsymbol{g}$). The vectorial geodesic dilation of size 1 of the marker image $\boldsymbol{g}$ with respect to the mask $\boldsymbol{f}$ can therefore be defined as:

$$\delta_{V}{}_{\boldsymbol{f}}^{(1)}(\boldsymbol{g}) = \delta_{V}{}^{(1)}(\boldsymbol{g}) \wedge_{V} \boldsymbol{f} \tag{4}$$

where $\delta_{V}{}^{(1)}(\boldsymbol{g})$ is the vectorial dilation of size 1 of the marker image *g*.

The vectorial geodesic dilation of size *n* of a marker color image **g** with respect to a mask color image ***f*** is obtained by performing *n* successive geodesic dilations of **g** with respect to ***f***:

$$\delta_{V}{}_{\boldsymbol{f}}^{(n)}(\boldsymbol{g}) = \delta_{V}{}_{\boldsymbol{f}}^{(1)}\left[\delta_{V}{}_{\boldsymbol{f}}^{(n\text{-}1)}(\boldsymbol{g})\right] \tag{5}$$

with $\delta_{V}{}_{\boldsymbol{f}}^{(0)}(\boldsymbol{g}) = \boldsymbol{f}$.

Geodesic transformations of images always converge after a finite number of iterations. The propagation of the marker image is limited by the mask image. Morphological reconstruction of a mask image is based on this principle [13].
The vectorial reconstruction by dilation of a mask color image ***f*** from a marker color image **g**, (both with D_f=D_g and $o_{lex}(\boldsymbol{g}) \leq o_{lex}(\boldsymbol{f})$) can be defined as:

$$R_{V}{}_{\boldsymbol{f}}(\boldsymbol{g}) = \delta_{V}{}_{\boldsymbol{f}}^{(n)}(\boldsymbol{g}) \tag{6}$$

where *n* is such that $\delta_{V}{}_{\boldsymbol{f}}^{(n)}(\boldsymbol{g}) = \delta_{V}{}_{\boldsymbol{f}}^{(n+1)}(\boldsymbol{g})$.

4 Highlight Detection by HLS, HSV and *L1-Norme*

Androutsos *et al* in [14] make a division of a luminance-saturation space and they conclude that if the saturation is greater than a 20% and the luminance is greater than a 75%, the pixels are chromatic, if the saturation is smaller than a 20% and the luminance is greater than 75%, the pixels are luminous or highlights. Our criterion is similar and it is based on the division of the LS, VS and MS diagrams in different homogenous regions that segment the pixels of the chromatic image. The exact limits of the regions must be calculated empirically, and in this comparative study we will show the exact values for the brightness region in the HLS, HSV and the *L1- norme.*

Not all the images have the same dynamic range and, therefore, the specularities do not present the same achromatic values and saturations. The best solution for this problem is to apply a vector morphological contrast enhancement for luminous pixels. We denote the color morphological contrast enhancement by:

$$\boldsymbol{f}' = \boldsymbol{f} + WTH_{V}(\boldsymbol{f}) \tag{7}$$

where ***f'*** is the new contrasted color image and WTH_v(***f***) is the vectorial top-hat (***f***-γ_v(***f***)) of the original color image ***f***. The color morphological contrast enhancement expels only the highlights to the limits of the RGB cube. The result of the local enhancement by the top-hat is that the specular reflectance pixels are located on c_2 line in LS and VS diagrams. In MS diagram the bright pixels are located on the c_3 and c_4 lines. The $s_{\max}$ value will be shown in the next section.

4.1 Use of LS, VS and MS Diagrams for Highlight Detection

We present the results of a study for highlight detection carried out on a set of real chromatic images that are quite representative of countless common materials (i.e., plastic, ceramics, fruit, wood, etc.), in which there are strong and weak reflectances. A subset of the images used in the study and its bi-dimensional diagrams (LS, VS, and MS) are present in Figure 2.

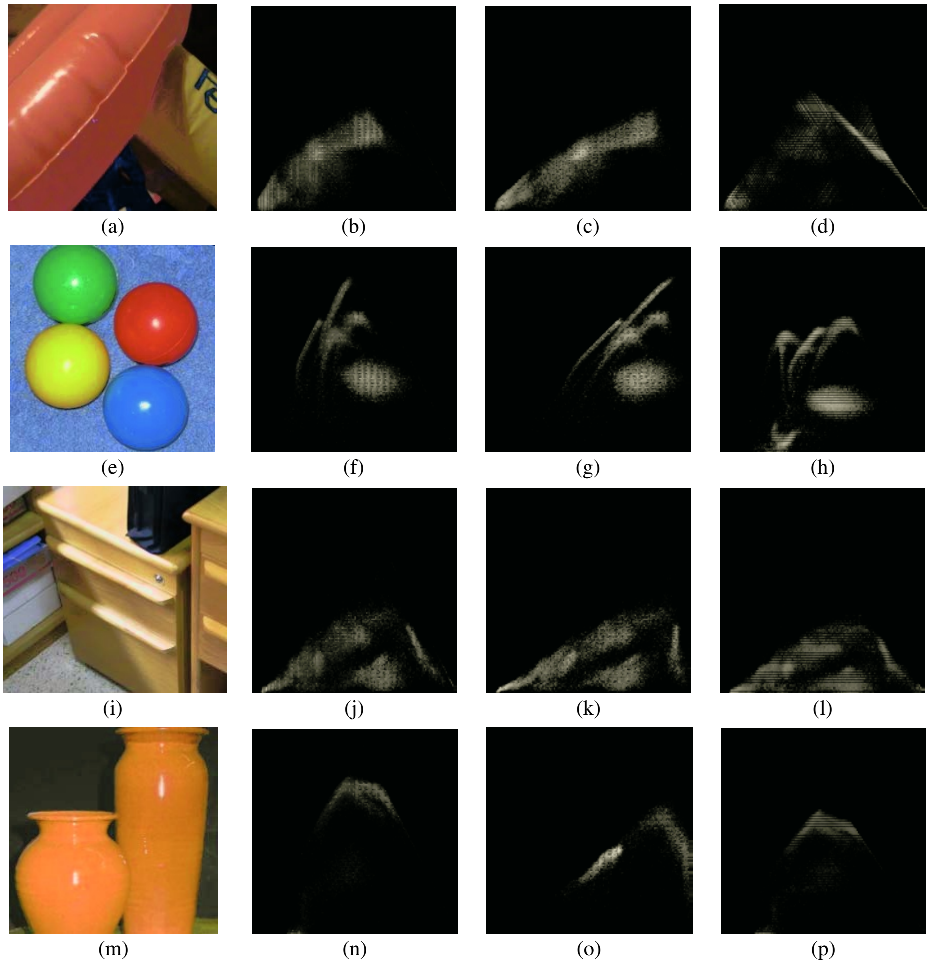
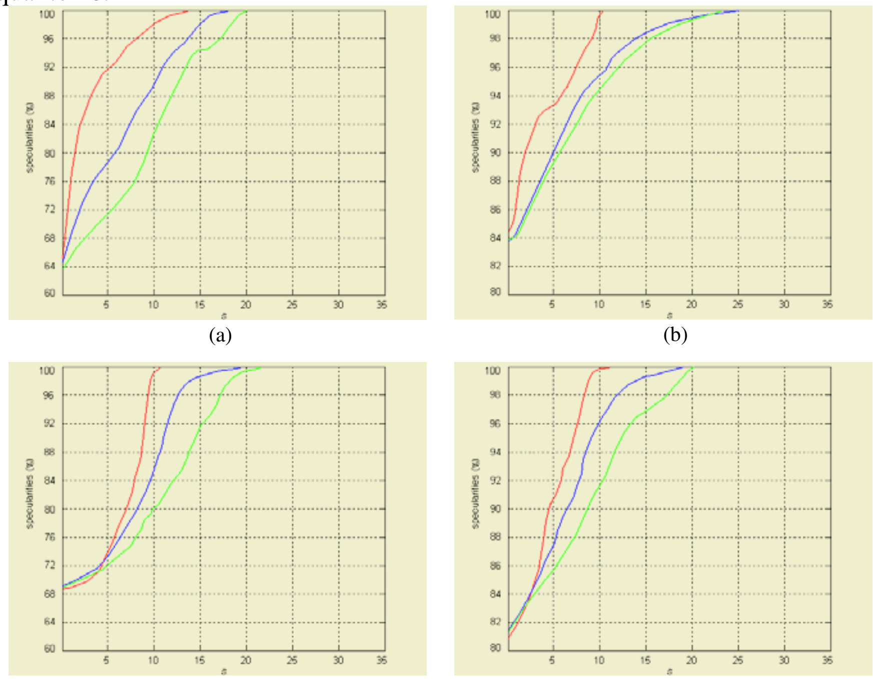


Fig. 2. Color images for empirical study and bi-dimensional histograms. "Life-saver" (a), "Balloons" (e), "Drawers" (i) and "Vases" (m). LS diagrams in (b), (f), (j) and (n). VS diagrams in (c), (g), (k) and (o). MS diagrams in (d), (h), (l) and (p).

Figure 3 shows the evolution of the highlights detected in Fig. 2 when the saturation s is increased along c_2 line (HLS and HSV), or c_3 and c_4 lines (MSH). It is a logarithmic evolution where most of the bright pixels are located in maximum value of achromatic signal and minimum *s* (up to 80%). The rest correspond to the transition from specular to diffuse reflection of the dichromatic refection model [1] in the surface of the objects. The graphs show that the detection of specularities stops, in all of the cases, at a maximum saturation of $s_{\max}$=25 (10% of 255). In HSV, $s_{\max}$ is smaller or equal to 18.



(a) (b)

(c) (d)

Fig. 3. Evolution (%) of highlights detected in bi-dimensional diagrams by saturation value *s*, in HSV (red line), HLS (blue line) and MSH (green line). (a) "Life-saver", (b) "Balloons", (c) "Drawers" and (d) "Vases".

5 Highlight Elimination by Vector Connected Filters

To eliminate the highlight that which was previously detected with the previous diagrams, we propose the use of geodesic filters of mathematical morphology [15]. Specifically, a vectorial opening by reconstruction applied exclusively to the specular areas of the image and their surroundings. In the filter, we use vector ordering $o_{\text{lex}}=m\rightarrow s\rightarrow h$ (*L1-norme*). A new mask-image $h(x,y)$ represents the pixels of ***f*** with which we will operate. The mask-image *h* is a dilation of the mask of highlights detected. Assuming $D_h=D_f$, each pixel (x,y) has a value of $h(x,y)=\{0,1\}$, where $h(x,y)=1$ in the new areas of interest in the image. The size *n* of the structural element of the dilation will determine the success of the reconstruction and the final cost of the operations, since this size defines the area to be processed by the filters. In the geodesic filter, ***f*** is first eroded with a structuring element of size *e*. The new filter is defined, considering that, in this case, the operation will not affect all the pixels (*x*,*y*), but only those in which *h*(*x*,*y*)=1:

$$R_{V\,\boldsymbol{f},h} = \left\{ \delta_{V\boldsymbol{f}}^{(n)}(\varepsilon_{V}^{(e)}(\boldsymbol{f})) \;\middle|\; \forall \boldsymbol{f}(x,y) \Rightarrow h(x,y)=1 \right\} \tag{8}$$

where *n* is such that $\delta_{V\mathbf{f}}^{(n)}(\varepsilon_{v}^{(e)}(\boldsymbol{f})) = \delta_{V\mathbf{f}}^{(n+1)}(\varepsilon_{v}^{(e)}(\boldsymbol{f}))$. If there is not brightness in the original image then, the reconstruction is not made.

From the results in Figure 4, the effectiveness of our method for the detection and elimination of specular reflectance can be observed. The over-simplification does not appear with the new filter, since the reconstruction only functions in bright areas. Furthermore, the results are obtained at a much lower computational time.

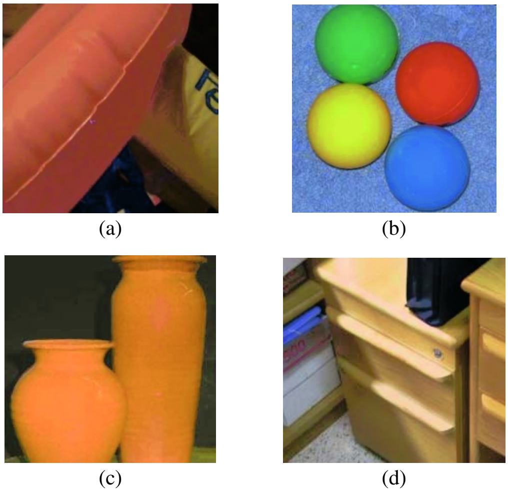


Fig. 4. Highlight elimination of real color images of Figure 2. Over-simplification is not present in the results.

6 Conclusions

In this paper, we have presented a comparative study about the detection and elimination of highlights in color images by color spaces based on polar co-ordinates. A detailed analysis has demonstrated that the brightness appear for different types of materials in a given area of the LS, VS and MS diagrams. The three spaces can be used, although the MSH allows the detection of brightness by different hue.

The use of a new connected vectorial filter allows us to eliminate the specular reflectance (highlights) previously detected. This filter is an extension of the geodesic transformations of the mathematical morphology to color images. The possibility of eliminating brightness in color images without causing over-simplification, has also been demonstrated. In addition, the elimination of brightness has been obtained automatically with a very short processing time. It is a reduction of temporal cost between 50% and 80%, with respect to a global geodesic reconstruction.

Based on the success shown by these results, the objective is to reduce the processing time required for geodesic operations as much as possible.

References

1. Shafer, S.A.: Using color to separate reflection components. Color Research Appl. Vol. 10 (1985) 210-218
2. Lin, S., Li, Y., Kang, S., et al: Diffuse-Specular Separation and Depth Recovery from Image Sequences. Lecture Notes in Computer Science, Springer-Verlag. Vol. 2352 (2002).
3. Bajcsy, R., Lee, S., Leonardis, A.: Detection of diffuse and specular interface reflections and inter-reflections by color image segmentation. International Journal on Computer Vision. Vol. 17 (1996) 241-271.
4. Klinker, G., Shafer, S.A., kanade, T.: Image segmentation and reflection analysis through color. In: Proc. SPIE. Vol. 937 (1988) 229-244.
5. Palus, H., Representations of colour images in different colour spaces. In: Sangwine, S., and Horne, R. (eds.): The Colour Image Processing Handbook, Chapman and Hall (1998) 67-90.
6. Plataniotis, K.N., Venetsanopoulos, A.N.: Color Image Processing and Applications, Springer-Verlag Berlin (2000)
7. Torres, F., Angulo, J., Ortiz, F.: Automatic detection of specular reflectance in colour images using the MS diagram. Lecture Notes in Computer Science, Springer-Verlag. Vol. 2756 (2003) 132-139.
8. Serra, J.: Espaces couleur et traitement d'images. Tech. Report N-34/02/MM. Centre de Morphologie Mathématique, École des Mines de Paris (2002)
9. Vicent, L.: Morphological Grayscale Reconstruction in Image Analysis: Applications and Efficient Algoritms. IEEE Transactions on Image Processing. Vol. 2. (1993) 176-201
10. Crespo, J. Serra, J., Schafer, R.: Theoretical aspects of morphological filters by reconstruction. Signal Processing. Vol. 47 (1995) 201-225
11. Ortiz, F., Torres, F., De Juan, E., Cuenca, N.: Colour mathematical morphology for neural image analysis. Journal of Real-Time Imaging. Vol. 8 (2002) 455-465
12. Ortiz F., Torres F., Gil P.: Gaussian noise elimination in colour images by vector-connected filters. In: Proc. IEEE 17th International Conference on Pattern Recognition. Vol. 4 (2004) 807-811
13. Soille, P.: Morphological Image Analysis. Principles and Applications. Springer-Verlag (1999)
14. Androutsos, D., Plataniotis, K., Venetsanopoulos, A.: A novel vector-based approach to color image retrieval using a vector angular-based distance measure. Computer Vision and Image Understanding, Vol. 75 (1999)
15. Ortiz F., Torres, F.: Vectorial morphological reconstruction for brightness elimination in colour images. Journal of Real-Time Imaging. Vol. 8 (2004) 379-387.

Algorithm for Crest Detection Based on Graph Contraction

Nazha Selmaoui

ERIM (Equipe de Recherche en Informatique et Mathématiques),
University of New Caledonia, B.P.4477, 98847 NOUMEA, New Caledonia
selmaoui@univ-nc.nc
http://pages.univ-nc.nc/~selmaoui/

Abstract. The concept of graph contraction was developed with the intention to structure and describe the image segmentation process. We consider this concept to describe a new technique of crest lines detection based on a definition of water-parting (or watershed). This technique allows to localize the basins delimited by these lines. The localization process uses the concept of ascending path. A structure of oriented graph is defined on original image. We give some definitions we use for this process. Before presenting the contraction algorithm, a pretreatment on the oriented original graph is necessary to make it complete. We show the algorithm resultson simple image examples.

1 Introduction

Graph Contraction (GC) is a theory that allows to built different types of hierarchies on top of such image graphs [2]. GC reduces the number of vertices and edges of image graph while, at the same time, the topological relations among the "surviving" components are preserved. The repeated process produces a stack of successively smaller graphs. This process is controlled by selected decimation parameters which consist of a subset of surviving vertices and associated contraction kernels [1].

The idea of the watershed ([4],[3]) is to attribute an **influence zone** to each of the regional minima of an image (connected plateau from which it is impossible to reach a point of lower gray level by an always descending path). The watershed is then defined as the boundaries of these influence zones. Numerous techniques have been proposed to compute the watershed. The major ones are reviewed in ([5], [6]). In one dimension, the location of the watershed is straightforward: it corresponds to the maxima of the function. In two dimensions (which is the case for gray level images), this characterization is not so easy. One can say in an informal way that the watershed is the set of **crest lines** of the image. The most efficient implementation described in the literature can be found in [6] and [7].

In this paper, we present a segmentation technique based on water-parting definition and developed by graph contraction process. The graph structure allows to take into account the image structure such as plateaus, step form etc.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 303–310, 2005.

In the first paragraph, we give briefly the contraction graph theory. In the second one, we give the watershed definition used here. We explain how to construct the graph structure, and how the contraction process will be applied according to the watershed definition. And finally, we perform this algorithm directly on an image in the aim to detect crest line, and also on image gradient to find edges.

$$\begin{tikzcd}
\text{neighborhood graph} & \text{face graph} \\
G_i(V_i, E_i) \arrow[r, leftrightarrow, "\text{dual}"] \arrow[d, Rightarrow, "\text{edge contracted}"'] \arrow[rd, phantom, "\textbf{dual edge contraction}\ \longrightarrow"] & \overline{G_i}(\overline{V_i}, \overline{E_i}) \arrow[d] \\
G^*(S_i, E_i \setminus N_{i,i+1}) \arrow[r, leftrightarrow, "\text{dual}"] \arrow[d, "\text{dually contracted}"'] \arrow[rd, phantom, "\longleftarrow\ \textbf{dual face contraction}"] & \overline{G^*}(\overline{V_i}, \overline{E_i} \setminus \overline{N_{i,i+1}}) \arrow[d, Rightarrow, "\text{eliminates } \deg(\overline{v}) < 3"] \\
G_{i+1}(V_{i+1}, E_{i+1}) \arrow[r, leftrightarrow, "\text{dual"] & \overline{G_{i+1}}(\overline{V_{i+1}}, \overline{E_{i+1}})
\end{tikzcd}$$

Fig. 1. Dual Graph Contraction: $(G_{i+1}, \overline{G_{i+1}}) = C[(G_i, \overline{G_i}), (S_i, N_{i,i+1})]$.

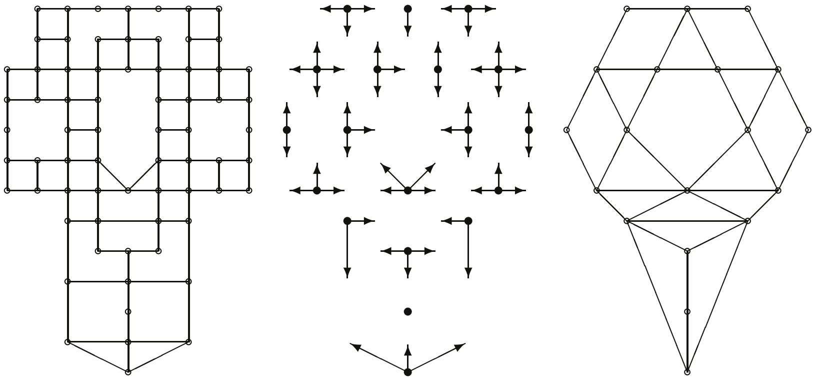
2 Graph Contraction

Dual graph contraction is the basic process [1] that builds an irregular 'graph' pyramid by successively contracting a dual image graph of one level into the smaller dual image graph of the next level. Dual image graphs are typically defined by the neighborhood relations of image pixels or by the adjacency relations of the region adjacency graph. The above concept has been used to find the structure of connected components [8]. Dual graph contraction proceeds in two basic steps (Fig. 1 in [1]): dual edge contraction and dual face contraction. The base of the pyramid consists of the pair of dual image graphs $(G_0, \overline{G_0})$. Following *decimation parameters* $(S_i, N_{i,i+1})$ determine the structure of an irregular pyramid [1]: a subset of *surviving vertices* $S_i = V_{i+1} \subset V_i$, and a subset of *primary non-surviving edges*[1] $N_{i,i+1} \subset E_i$. Every non-surviving vertex, $v \in V_i \setminus S_i$, must be connected to one surviving vertex in an unique way. The relation between the two pairs of dual graphs, $(G_i, \overline{G_i})$ and $(G_{i+1}, \overline{G_{i+1}})$, as established by dual graph contraction with decimation parameters $(S_i, N_{i,i+1})$ is expressed by function $C[.,.]$:

$$(G_{i+1}, \overline{G_{i+1}}) = C[(G_i, \overline{G_i}), (S_i, N_{i,i+1})] \tag{1}$$

The contraction of a primary non-surviving edge consists in the identification of its endpoints and in the removal of both contracted edge and its dual edge. Dual face contraction simplifies most of the multiple edges and self-loops, but not those inclosing any surviving parts of the graph (see [1]). One step of dual

[1] Secondary non-surviving edges are removed during dual face contraction.



(a) (V_0, E_0) (b) $(S_0, N_{0,1})$ (c) (V_1, E_1)

Fig. 2. Example of dual graph contraction: $(V_1, E_1) = C[(G_0, \overline{G_0}), (S_0, N_{0,1})]$.

graph contraction is illustrated in Fig. 2. Note that the contracted graph may contain both self-loop and multiple edges. They are necessary to preserve the structure defined in the base graph [1]. Decimation parameters control dual graph contraction, a process that iteratively builds an irregular (graph) pyramid.

3 Image Segmentation Using GC Process

In this paragraph, we present the segmentation algorithm. For this, we choose an appropriate watershed definition which fit with GC process. A structure of oriented graph must be defined in digital image.

3.1 Definitions and Principle

There are many definitions of watershed, it seems easy to define them on digital pictures. However, when looking closer at these definitions, it turns out that there are many particular cases. Let us consider a two-dimensional grayscale picture f whose definition domain is denoted $D \subset Z^2$. f is supposed to take discrete (gray) values in a given range $[0, N]$, N being an arbitrary positive integer:

$$f \begin{cases} V \subset Z^2 \rightarrow \{0, 1, ..., N\} \\ v \qquad\quad \mapsto f(v) \end{cases} \tag{2}$$

Let G denote the graph of underlying digital grid, which can be of any type: a square grid in four or eight connectivity, or a hexagonal grid in six connectivity. G is a subset of $Z^2 \times Z^2$.

Definition 1. *A path P of length l between two pixels v and w in a picture is a $(l+1)$-tuple of pixels $(v_0, v_1, ..., v_{l-1}, v_l)$ such that $v_0 = v, v_l = w$, and $\forall i \in [1, l], (v_{i-1}, v_i) \in G$.*

In the following, we denote $l(v)$ the length of a given path P. We also denote $N_G(v)$ the set of the neighbors of a pixel v, with respect to G: $N_G(v) = \{w \in Z^2, (v,w) \in G\}$.

Definition 2. *A minimum M of f is a connected plateau of pixels from which it is possible to reach a point of lower altitude without having to climb:*

$$\forall v \in V, \forall w \notin M, \textit{ such that } f(v) \leq f(v),$$

$$\forall P = (v_0, ..., v_l) \textit{ such that } v_0 = v \textit{ and } v_l = w,$$

$$\exists i \in [1, l] \textit{ such that } f(v_i) > f(v) \tag{3}$$

A minimum is connected area where the gray level is strictly less than on the neighboring pixels level. We can now give the definition of watershed that we are using here in term of graph.

Definition 3. *(In term of steepest slope lines): The catchment basin C(M) associated with a minimum M is the set of pixels v of V_f such that a water drop falling at v flows along the relief, following a certain descending path called the downstream of v [9], and eventually reaches M.*

The lines which separate different catchment basins, build the watershed (or dividing lines). We can formulate this definition in term of graph and it allows to compute the decimation parameters which control the GC process such that pixels of crest survive.

In according to definitions 2 et 3 and the descending path, we must define the graph where the paths are oriented. Therefore, we transform the initial graph G to a directed graph $G_d = (V_d, E_d)$ in the following way:

Definition 4. *$G_d(V_d, E_d)$ is the directed graph such that $V_d = V$ and E_d defined by:*

$$\forall v_i, v_j \in V_d, e = (v_i, v_j) \in E_d \Leftrightarrow \begin{cases} e = (v_i, v_j) \in E \\ f(v_j) = Sup\{f(v_k) \mid f(v_i) < f(v_k), \forall (v_i, v_k) \in E\} \end{cases} \tag{4}$$

The notion of oriented path can be defined as following:

Definition 5. *An oriented path $\boldsymbol{P} = (v_0, ..., v_l)$ is a path such that there exist an ascending path from v_0 to v_l i.e.: $\forall i \in [1, l], (v_{l-1}, v_l) \in E_d$.*

Figure 3 illustrates graph G_d on a simple example. As shown in figure 3, the graph is not complete, there are not connected pixels of plateau. This is due to a strictly inequality in the oriented graph definition G_d, and a large inequality generates “circuits”. Pixels which are not the terminal endpoint and not the initial endpoint belong to zones of plateau. We will explain how to connect pixels of plateau.

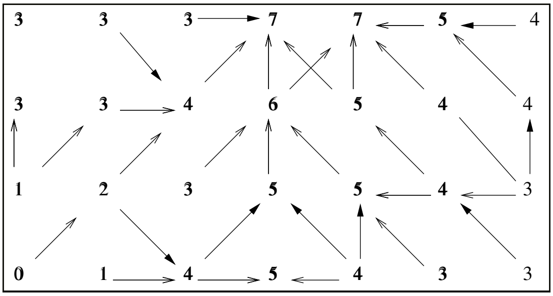


Fig. 3. Directed graph on an image example.

3.2 Connection of Plateaus

To connect the plateau's pixels, we introduce the new edges in the graph. Then we obtain a new graph G_d^* containing all graph's edges, and the new edges verifying the following definition:

Definition 6.

$$e = (v, w) \in E_d^* \Leftrightarrow \begin{cases} w \ is \ not \ a \ final \ endpoint \ in \ G_d \\ f(v) = f(w) \end{cases} \tag{5}$$

As shown in figure 4a representing the graph G_d^*, "circuits" appear between plateau's pixels. We will proceed to a preliminary pre-processing before final contraction. This allows to contract all pair of pixels connected with two edges to one node of graph (see figure 5a).

In spite of plateau's connection, some pixels on ascending path don't belong to edges as terminal endpoint. We need to add other edges such as every pixel must be at least a terminal endpoint for one edge. Graph G_d^* changes into graph G_d^{**} with the following conditions:

$$\forall v \in V_d^* \text{ such that } \nexists (w, v) \in E_d^* \text{ then } (z, v) \in E_f^{**}$$

where

$$f(z) = Inf\{f(v_i) \ / \ f(v) \geq f(v_i), \forall (v_i, v) \in N_G(v)\}.$$

Figure 4b shows the graph G_d^{**}.

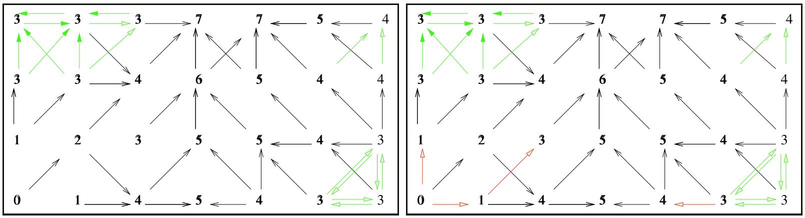
3.3 Decimation Parameters and Algorithm

The formal definition allows to compute the decimation parameters : *Every node v which is the unique terminal endpoint of an edge $e = (w, v)$ belongs to the same zone as initial endpoint w of e.*

Definition 7. *The decimation parameters (S, N) for GC to watershed are:*

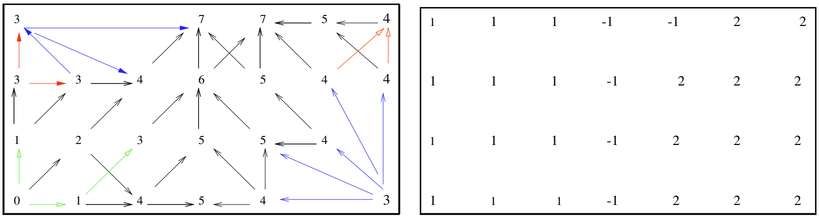
$$NS = V_d^{**} \backslash S$$

$$where \ S = \{v \in V_d^{**} / \ \nexists e \in E_d^{**}; v = terminal - endpoint(e)\}$$
$$N = \{e \in E_d^{**} / \exists v \in S; v = terminal - endpoint(e)\} \tag{6}$$



(a) connection of plateau (b) A complete graph

Fig. 4. Preprocessing of graph.



1	1	1	-1	-1	2	2
1	1	1	-1	2	2	2
1	1	1	-1	2	2	2
1	1	1	-1	2	2	2

(a) Contraction of regional minima (b) Final result after graph contraction

Fig. 5. Graph contraction process.

The algorithm is given below:

- Transformation of image to graph G_d
- Processing of plateaus
- Contraction of plateaus
- repeat
  - Compute the decimation parameters (S, N)
  - Graph contraction
- until $S = \emptyset$

After last contraction, the final graph content two types of nodes, the first one presenting the basins, and the others presenting the water-parting line. As shown in figure 5b, pixels with value 1 and 2 constitute two basins, pixels with value -1 constitute the water-parting line.

4 Results and Conclusion

Two images are used to test our algorithm, one image gradient for edge detection (figure 6b) and one raw image for crest detection (figure 6a). The obtained results (figure 7) on image figure 6a are satisfactory, because no post-processing has been necessary. On the other hand, an over-segmentation has been observed. Thresholding crest with depth less than a fixed threshold by user has been realized. Consequently, this type of algorithm is well adapted to images containing lines than images with edges.

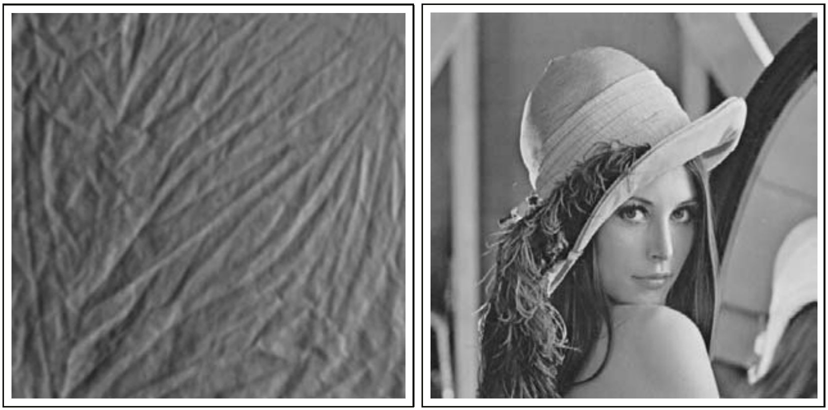
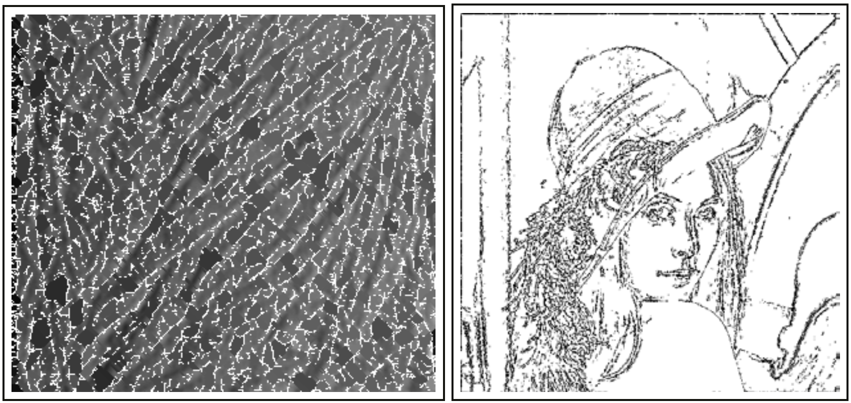


Fig. 6. Original images.



(a) crest line detection

(b) edge detection

Fig. 7. Image results.

References

1. Kropatsch, Walter G.: Building Irregular Pyramids by Dual Graph Contraction. IEE-Proc. Vision,Image and Signal Processing, Vol.142(No.6), (1995) 336–348
2. Kropatsch, Walter G. and Ben Yacoub, S.: Universal Segmentation with PIRRAMIDS. 20th Workshop of the Austrian Association for Pattern Recognition (ÖAGM/AAPR), (1996) 171–182
3. Beucher, S. and Lantuéjoul, C.: Use of Watersheds in Contour Detection. Int'l Workshop Image Processing, Real-Time Edge and Motion Detection/Estimation, Rennes, Paris, 1979 17-21
4. Digabel, H. and Lantuéjoul, C.: Iterative Algorithms. Second European Symp. Quantitaive Analysis of Microstructures in Material Science, Biology and Medicine, Stuttgart, Germany: Rieder, (1978), 85–99

5. Vincent, L.: Algorithmes Morpholgiques á base de Files d'Attente et de Lacets: Extension aux Graphes. Thése de doctorat, Ecole des Mines, Paris, France, (1990).
6. Vincent, L. and Soille, P.: Wtersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations. IEEE Trans. Pattern Analysis and Machine Intelligence Vol.13 (No6), (1991), 583–598
7. Soille, P.: Morphological Image Analysis: principles and Applications. Springer-Verlag, Berlin, Heidelberg, New York. Press, (1999)
8. Macho, H. and Kropatsch, W.G.: Finding Connected Components with Dual Irregular Pyramids. Proc. of 19th ÖAGM and 1st SDVR Workshop. Solina, F. and Kropatsch, W.G. (1995),313–321.
9. Beucher, S.: Segmentation d'images et morphologie mathématique. Thèse Doctorat: Ecole Nationale Supérieure des Mines de Paris. (1990), 295p.

A Learning Framework for Object Recognition on Image Understanding

Xavier Muñoz, Anna Bosch, Joan Martí, and Joan Espunya

Institute of Informatics and Applications, University of Girona,
Campus de Montilivi s/n 17071 Girona, Spain
{xmunoz,aboschr,joanm,jespunya}@eia.udg.es

Abstract. In this paper an object learning system for image understanding is proposed. The knowledge acquisition system is designed as a supervised learning task, which emphasises the role of the user as teacher of the system and allows to obtain the object description as well as to select the best recognition strategy for each specific object. From several representative examples in training images, an object description is acquired by considering different model representations. Moreover, different recognition strategies are built and applied to obtain initial results. Next, teacher evaluates these results and the system automatically selects the specific strategy which best recognise each object. Experimental results are shown and discussed.

1 Introduction

The aim of image understanding systems is easy to describe. Given an arbitrary photograph, we would like to automatically understand and give meaning to the image by identifying and labeling significant objects in the image. Nevertheless, although we human perform this perception in an immediate and effortless manner, to adequately describe a scene significantly involves the integration of different image processing techniques, pattern recognition algorithms, and artificial intelligence tools, and is a very difficult problem in computer vision [1].

An image understanding system can also be considered as a knowledge-based vision system, because such system requires models that represent prototype objects and scenes. Hence, two important issues must be taken into account: (1) the way in which the model knowledge is organized and stored, and (2) how this knowledge is acquired. However, while knowledge representation has become a permanent focus of interest and a large number of proposals can be found in the literature (see [2]), knowledge acquisition tools are still in their infancy [2].

Early systems were generally not oriented to facilitate the entry of knowledge or carry out some form of automated learning. In contrast, most of the existing systems had to incorporate this new model knowledge by hand; and all the more so for code-encapsulated data. Examples are the Schema system [3] and the region analyzer of Ohta et al. [4], which are successful systems and works of reference. Nevertheless, nowadays most vision researchers agree that the success of scene description systems lies on their ability to learn from experience and

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 311–318, 2005.

training, and there is in last years a clear trend towards the consideration of learning as one of main issues that need to be tackled in designing a visual system for object recognition [5, 6].

Automated learning must consider the acquisition of object models as a description of the object attributes as well as a selection of the strategy used to find and recognize it in an image. Actually, not all objects are defined in terms of the same attributes, and even these attributes may be used in various ways within the matching or interpretation process. Therefore, the learning system must take a flexible and multifaceted recognition strategy into account. A large number of object recognition strategies have been proposed to achieve a particular goal. However, we think is too much pretentious to think that a single method will be able to correctly model and recognize all objects in the real world visual gallery. There is not a perfect strategy for all objects and very little research in the field of computer vision has gone into the problem of determining the best recognition strategies [7].

In this paper we propose an object learning framework for image understanding mainly oriented to outdoor scene images, which addresses the problem of automatic object recognition strategy selection. Inspired on relevance feedback techniques used on image retrieval systems [8], the knowledge acquisition system is designed as a supervised learning task, which involves the user as teacher and part of the learning process. Therefore, the learning allows to obtain the object description as well as to select the best recognition strategy for each specific object under a friendly and effortless user interface.

The remainder of this paper is structured as follows: Section 2 describes the proposed supervised object knowledge learning, focusing on the object description, recognition strategy design and the strategy selection. Experimental results proving the validity of our proposal appear in Section 3. Finally, conclusions are given in Section 4.

2 Learning Proposal

Models constitute representations of the real world, and thus, modeling implies the choice of a suitable set of parameters in order to build those representations. Obviously, the selection of the parameters will affect how well the models fit reality, and this becomes a central issue in any modeling process. Due to the complexity of outdoor scenes, our approach includes the possibility that every single object can be described by specific features and a specific recognition strategy, facilitating later recognition processes. With the aim to provide some improvements in knowledge engineering tasks, system code and models databases become totally independent, with a fast, simple and easy data acquisition process.

Our object learning approach has been designed as a supervised task, which emphasizes the role of the user as the responsible of teaching the system. First, the teacher selects representative examples of objects in training images. From these examples, an object description is acquired by considering different model

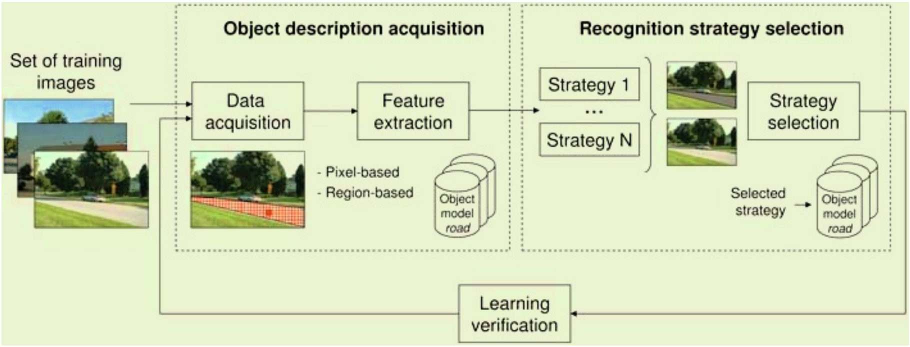


Fig. 1. Scheme of the proposed knowledge acquisition process, working as a supervised learning task.

representations. Next, several strategies to recognize the object are built and applied in order to obtain initial recognition results. These results are evaluated by the teacher, and the system automatically selects the specific strategy which best recognize each object. A global scheme of the proposed knowledge acquisition process is shown in Figure 1.

2.1 Object Description Acquisition

The teacher firstly selects meaningful examples of objects in the training images by clicking on a pixel corresponding to the object of interest. This simple selection allows us to extract the whole object and to compute and register different model representations which provide a complete description of the object. Specifically, the acquired information is composed by:

- **Pixel-based description:** from the selected point, a set of neighboring pixels are extracted and considered as samples of the object pixels. Next, a large number of color and texture features of these pixels are measured. We initially consider the whole set of features as candidates to characterize real objects. In particular, 28 color features related to different color spaces and a set of 8 co-occurrence matrix based texture features are computed for each pixel.
- **Region-based description:** a color texture active region which integrates region and boundary information [9] grows from the selected point in order to segment the region corresponding to the whole object. Color and texture descriptors, as well as shape information based on Fourier descriptors [10], are extracted for each region in order to describe and characterize the object of interest.

2.2 Object Recognition Strategies

The object descriptions can be used in different ways in order to recognize the object. We initially selected, implemented and included into our framework three simple and basic recognition algorithms, which differ on the description model they use as well as the philosophy of the strategy. Recognition strategy A is a top-down approach based on the pixel-level description; recognition strategy B follows a classic bottom-up approach based on a general non-purpose segmentation; while strategy C is a pure hybrid strategy which applies a knowledge-guided search over an initial segmentation. More specifically,

Recognition Strategy A. The top-down strategy consists on the direct search of a specific object by exploiting information concerning the object's characteristics. The implemented method, similarly to the proposal of Dubuisson-Jolly and Gupta [11], models the different objects by a multivariate Gaussian distribution. A pixel-level classification is obtained by using the maximum likelihood classification technique which assigns each pixel to the most probable object.

Recognition Strategy B. A classical bottom-up approach for image understanding was considered for this method. The technique is mainly based on a general purpose segmentation step which tries to part the image into meaningful regions. A color texture segmentation algorithm based on active regions, which integrates region and boundary information [9] was used for this purpose on our implementation. Next, main regions are labeled according to their similarity with stored models.

Recognition Strategy C. Finally, the last implemented method can be considered as a pure hybrid strategy, which starts as the previous approach with a general segmentation. However, a top-down strategy is then performed over these results to specifically find objects of interest. As was noted by Draper et al. [7], not all object classes are defined in terms of the same attributes, and a previous feature selection process allows to select the specific subset of features which best characterizes each single object. Next, selected features are considered to look for the segmented regions on the image which match with the object model.

2.3 Recognition Strategy Selection

Once the process of recognition strategies design is complete, the best specific strategy to recognize each object must be determined. This is the key stage of our proposal; inspired on relevance feedback techniques extensively used on content-based image retrieval systems [8], the role of the user is emphasized and he/she is involved as a vital part of the learning process. With the help of the teacher interaction, the system is able to evaluate the different recognition strategies and to learn which is the best strategy for each object.

Therefore, given a reduced set of training images, the recognition methods are launched together to find all the instances of the given object. Obviously, these strategies can provide different results: because a strategy misses an object apparition, or contrarily it gives a false positive. And in all cases the extraction accuracy must be determined. Hence, these recognition results are visually retrieved to the teacher in order to evaluate their quality. In front of these results, the teacher marks the found instances to indicate if they are well recognized or not. In other words, if the strategy (or strategies) which obtained this recognition was right or wrong. We provide the teacher with three levels of correctness: highly correct, correct and wrong. Although the use of more levels could probably provide more information, we consider it would be lesser friendly for the user to interact with the system. When results have been evaluated by the teacher, this information allows to the system measure the score of each strategy and finally select the strategy which best recognize the object.

The learning process ends with a final verification step. A visual feedback is provided by means of recognizing the object in the set of training images. Obviously, the specific selected strategy will be used for each object. This visual feedback guides the teacher, giving him the option to interfere in the learning process by introducing new training images.

3 Experimental Results

We applied our method to a color image data set constructed using 100 images from the image database of the University of Oulu [12] and also a set of images taken by ourself. These images consists on natural outdoor scenes and mainly contain typical objects in rural and suburban area. We segmented and labeled them manually into 4 classes: *sky*, *grass*, *trees*, *road*, while the remaining areas are considered as *unknown* objects. The training set includes 20 selected images and the remaining 80 were used for testing. We evaluate both, the method selection from the user interaction (section 3.1), and the final goodness of the recognition using the selected strategy (section 3.2). Furthermore, the system is available on an on-line web-based application at *http://ryu.udg.es*.

3.1 Learning Results

The selection of a specific object recognition method from the user feedback and how the system is able to capture the user's criterion, is evaluated by measuring its match with the quantitative results obtained for the different techniques over the training images. Ideally, the selected method must be that which achieves the best results. Table 1 summarizes the scores assigned to each method from the teacher judgements. For example, the user qualifies the top-down strategy for recognizing the road with a quality percentage of 87.50%, which means that the user mostly agrees with the results obtained by this strategy in the recognition of the *road* object. The method which obtains the best score is specifically selected for each object. Summarizing, the top-down strategy A was selected for the recognition of *road* and *grass*, while the hybrid strategy C was considered for the *sky* and *trees*.

On the other hand, Table 2 shows the quantitative evaluation of the strategies over the same training images by measuring the percentage of well classified and over-classified pixels. As was desirable, the strategies selected by the user to classify each object achieve the best results, which means the system is able to capture the user feedback and selects the best method over the set of training images. However, the selection for the *sky* object must be explained. In this case, the system selected the strategy C while the strategy A obtained a 100% of well classified pixels. Nevertheless, this initially wrong decision can be justified by the high percentage of wrongly over-classified pixels which obtains the A top-down strategy. Figure 2 shows the object classification obtained by both techniques over some images of the training set. As is stated in the first column, the top-down strategy A correctly extracts the sky, but confuses some road pixels with this object, while a better recognition is achieved by the hybrid strategy C, which reaffirms the selection performed by our system from the user feedback.

Table 1. Scores obtained to recognize each object taken into account the user criterion. (TD = Top-down; BU = Bottom-up; H = Hybrid.)

Percentage acquired from the user				
Strategy	sky	road	grass	trees
strategy A (TD)	37.50%	87.50%	50.00%	−50.00%
strategy B (BU)	10.00%	20.35%	24.5%	10.18%
strategy C (H)	100.00%	71.43%	35.00%	44.12%

Table 2. Percentages of well classified and over-classified pixels over the training images. (TD = Top-down; BU = Bottom-up; H = Hybrid.)

Percentage acquired from the training classified images								
Strategy	sky		road		grass		trees	
	ok	over	ok	over	ok	over	ok	over
strategy A (TD)	100.00%	10.01%	89.24%	0.25%	87.55%	1.39%	73.36%	9.31%
strategy B (BU)	60.00%	15.00%	60.49%	3.00%	61.04%	7.00%	63.87%	5.00%
strategy C (H)	90.15%	4.38%	75.84%	0.43%	83.13%	3.65%	88.31%	3.13%

3.2 Classification Results

Table 3 summarizes the object recognition results obtained by both selected strategies (top-down strategy A and hybrid strategy C) over the test images set. Moreover, the last row show the percentages obtained by each selected specific object method. The last columns shows the percentages taken into account all objects. From these quantitative results, the significant improvement that is achieved by the use of a specific method for each single object and, the combination of strategies on the whole system in front of a single technique, is stated. Specifically, using the set of selected strategies the system obtains a 85.30% of well-classified pixels, which is clearly superior to the scores obtained by both individual techniques. Moreover, the lowest percentage of wrongly over-classified pixels, 1.88%, is obtained. The results can be qualitatively evaluated in Figure 3,

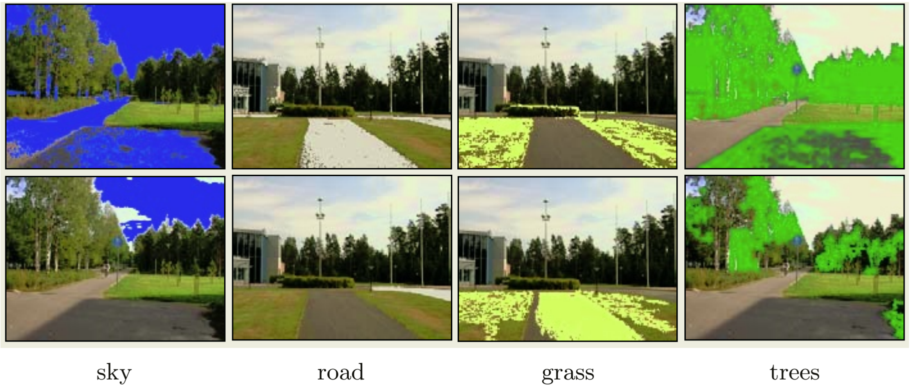


Fig. 2. Some labeling results over the training images set. First row shows results by the top-down strategy A; while second row shows results by the hybrid strategy C.

Table 3. Object classification and over-classification rates over the test images. The last row shows the percentages achieved by the object specific strategy selected from the user feedback.

Percentage acquired from the testing classified images										
Strategy	sky		road		grass		trees		average	
	ok	over	ok	over	ok	over	ok	over	ok	over
strategy A	64.49%	4.94%	86.46%	0.30%	83.20%	1.38%	67.79%	10.69%	75.48%	4.30%
strategy C	83.92%	0.37%	67.24%	0.33%	46.06%	2.27%	87.62%	5.51%	71.21%	2.80%
selected	83.92%	3.15%	86.46%	0.30%	83.20%	1.38%	87.62%	5.51%	85.30%	1.88%

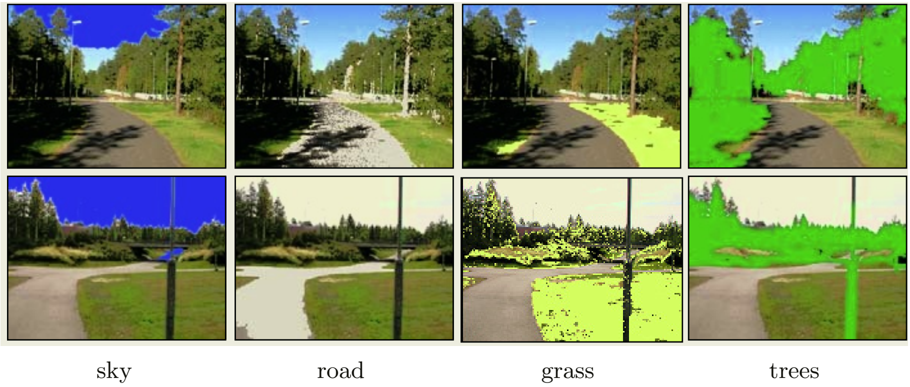


Fig. 3. Some labeling results over the test images set using by each object the strategy selected by the proposed learning framework.

which shows the object recognition achieved by our system using specifically selected object recognition techniques, and denotes the correctness of the object learning and extraction.

4 Conclusions

An object learning framework for image understanding has been described. The process has been designed as a supervised learning task, which emphasizes the role of the user as system teacher. From some examples provided by the teacher, the system extracts the information required to describe the object. Moreover, the learning allows to select the best recognition strategy for each specific object under a friendly and effortless user interface. Experimental results has stated the convenience of using a set of object specific recognition methods.

Extensions of this work are oriented to the improvement of the strategy selection to make possible the combination of several techniques as the best method to recognize an object. Furthermore, new recognition strategies will be included into the system.

References

1. Yun-tao, Q.: Image interpretation with fuzzy-graph based genetic algorithm. In: IEEE International Conference on Image Processing, Kobe, Japan (1999) 545–549
2. Crevier, D., Lepage, R.: Knowledge-based image understanding systems: A survey. Computer Vision and Image Understanding **67** (1997) 161–185
3. Draper, B., Collins, R., Brolio, J., Hanson, A., Riseman, E.: The schema system. International Journal of Computer Vision **2** (1989) 209–250
4. Ohta, Y.: Knowledge-based Interpretation of Outdoor Natural Color Scenes. Pitman Publishing, Boston, Massachussets (1985)
5. Drummond, T.: Learning task-specific object recognition and scene understanding. Computer Vision and Image Understanding **80** (2000) 315–348
6. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Volume 2. (2003) 264–271
7. Draper, B., Hanson, A., Riseman, E.: Knowledge-directed vision: Control, learning, and integration. Proceedings of the IEEE **84** (1996) 1625–1637
8. Rui, Y., Huang, T.S. Ortega, M., Mehrotra, S.: Relevance feddback: A power tool for interactive content-based image retrieval. IEEE Trans on Circuits and Systems for Video Technology **8** (1998) 644–655
9. Muñoz, X., Freixenet, J., Cufí, X., Martí, J.: Active regions for colour texture segmentation integrating region and boundary information. In: IEEE International Conference on Image Processing, Barcelona, Spain (2003)
10. Zhang, D., Lu, G.: Generic fourier descriptor for shape-based image retrieval. In: IEEE International Conference on Multimedia and Expo. (2002)
11. Dubuisson-Jolly, M.P., Gupta, A.: Color and texture fusion: Application to aerial image segmentation and gis updating. Image and Vision Computing **18** (2000) 823–832
12. Ojala, T., Mäenpää, T., Pietikäinen, M., Viertola, J., Kyllönen, J., Huovinen, S.: Outex - new framework for empirical evaluation of texture analysis algorithms. In: IAPR International Conference on Pattern Recognition. Volume 1. (2002) 701–706

A Roof Edge Detection Model

Qing H. Zhang, Song Gao, and Tien D. Bui

Department of Computer Science and Software Engineering
Concordia University, Montreal, QC, Canada

Abstract. We have generalized the Mumford-Shah model to obtain a new model capable of detecting roof edges. In this new model, we have assumed a piecewise planar surface for each bounded region. We have also shown that this new model is less dependent on the scale parameter ν than the original Mumford-Shah model. We have proved that the gradient projection method can produce the minimal energy. The validity of the new model is demonstrated by experimental results.

1 Introduction

The segmentation problem is an important topic in many different areas including computer vision, medical imaging, video processing etc. This problem can be defined as follows: For an observed image u_0 (possibly with noise), we want to find an optimal piecewise smooth approximation u of u_0 for each specific region. The regions are denoted by $\Omega_i, i = 1, 2, ..., n$. The function u varies smoothly within each Ω_i, and rapidly or discontinuously across the boundaries of Ω_i. The process of finding the boundaries of Ω_i is called segmentation. The boundaries of all Ω_i are denoted as C and Ω_i is an open set. Therefore the whole image can be expressed as $\Omega = \bigcup \Omega_i \bigcup C$.

To find the boundaries, Mumford and Shah have proposed the following minimization problem [1, 8]:

$$inf_{u,C}\{E(u, C) = \alpha \int_{\Omega \backslash C} (u - u_0)^2 dxdy + \mu \int_{\Omega \backslash C} |\nabla u|^2 dxdy + \nu|C|\}, \quad (1)$$

where $\mu, \nu, \alpha > 0$ are parameters which can be considered as weight factors. The problem is to find u and C such that the above energy is minimal. C is the segmentation curve and u is the approximation of u_0.

From Eq.(1), we have the following observations [8]: For the value of $E(u, C)$ to be small, we need (1) u is a good approximation of u_0 (2) u does not vary much in each region Ω_i and (3) the boundary of each region Ω_i is as short as possible. The minimal value of $E(u, C)$ depends on the values of α, μ and ν.

2 Markov Random Field Model

The Mumford-Shah (MS) model can be understood in the following way: For an observed noisy image u_0, we would like to find the true (clean) image u. Using

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 319–327, 2005.

Bayesian decision theorem, the posterior probability is $P(u|u_0) = P(u_0|u)P(u)$. Here $P(u)$ is the probability of obtaining u, and $P(u_0|u)$ is the probability of obtaining u_0 given the image u. Then $P(u|u_0)$ is the probability of obtaining the (clean) segmentation image u given u_0. We assume that $P(u)$ is proportional to the energy in the following form [16] $P(u) = \exp(-U)$. Here U is energy of the image. Considering only interactions between neighboring site of the image, U can be expressed as

$$U \propto \sum_{i\ and\ j\ are\ neighbor\ pixels} g(u(i) - u(j)) \propto \sum g(\nabla u). \tag{2}$$

Where $g(x)$ is the regularization function; i and j are the indices of pixels in the image. $P(u_0|u)$ is assumed to be

$$P(u_0|u) \propto \prod_{i=1}^{N} \exp[-(u_0(i) - u(i))^2] \propto \exp[-\int (u_0 - u)^2 dxdy]. \tag{3}$$

Here i is the index of a pixel and N is the total number of pixels in the image. Thus, $P(u_0|u)$ increases as u approaches u_0. Assuming $P(u|u_0) = \exp(-E(u, u_0))$, then we have

$$E(u, u_0) = \int (u - u_0)^2 dxdy + \mu \int g(\nabla u) dxdy. \tag{4}$$

If we include the boundary energy in Eq.(4), we have

$$E(u, u_0) = \int (u - u_0)^2 dxdy + \mu \int g(\nabla u) dxdy + \nu |C|. \tag{5}$$

This is the MS model if we take $g(x) = x^2$.

Finding the solution of Eq.(1) for an arbitrary image is not a trivial task [17]. Therefore, many approximations and simplified models have been proposed [2, 4–6, 12–14, 18]. For the segmentation problem, the boundary is most important. The texture inside the boundaries is secondary. As the first approximation, we can take u as a constant in each region. This is called piecewise constant model and is widely used in image segmentation [2].

3 Piecewise Constant Approximation

If we assume that u is a constant in each region ($u = c$), Eq.(1) becomes to

$$inf_{u,C}\{E(u, C) = \alpha \int_{\Omega \setminus C} (c - u_0)^2 dxdy + \nu |C|\} \tag{6}$$

A simplified version of Eq.(6) has been studied in detail by Chan and Vese [2] by using the level set method. In the lelvel set formulation, the curve C is replaced by the level set function $\phi(x, y, t)$

$$\phi(x, y, t) = \begin{cases} > 0 & \text{if} \quad (x, y)\ in\ \Omega \\ = 0 & \text{if} \quad (x, y)\ in\ \delta\Omega \\ < 0 & \text{if} \quad (x, y)\ in\ \bar{\Omega} \end{cases}$$

With the help of the Heaviside function $H(x)$, the two-phase version of Eq.(6) can be written as

$$E(c_1, c_2, \phi) = \int F \; dxdy \tag{7}$$

where $F = \alpha(c_1 - u_0)^2 H(\phi) + \alpha(c_2 - u_0^2)(1 - H(\phi)) + \nu\delta(\phi)|\nabla\phi|$, c_1 and c_2 are constants.

Since c_1 and c_2 are constants, the minimization of Eq.(7) can be written as:

$$inf_\phi\{E(\phi) = \int_\Omega F(x, y, \phi, \phi_x, \phi_y) \; dxdy\}, \tag{8}$$

with $\phi = \phi(x, y, t), \phi_x = \frac{\partial\phi(x,y,t)}{\partial x}, \phi_y = \frac{\partial\phi(x,y,t)}{\partial y}$, we obtain the following Euler-Lagrange equation

$$\frac{\partial F}{\partial\phi} - \frac{\partial}{\partial x}(\frac{\partial F}{\partial\phi_x}) - \frac{\partial}{\partial y}(\frac{\partial F}{\partial\phi_y}) = 0 \tag{9}$$

with boundary condition $\frac{\partial F}{\partial\phi_x}\frac{\partial x}{\partial\hat{n}} + \frac{\partial F}{\partial\phi_y}\frac{\partial y}{\partial\hat{n}} = 0$. Here $\hat{n}$ is the normalized normal to $\delta\Omega$. Substituting F from Eq.(7) into Eq.(9), we have

$$\delta(\phi)[(c_1 - u_0)^2 - (c_2 - u_0)^2 - \nu\nabla\cdot(\frac{\nabla\phi}{|\nabla\phi|})] = 0 \tag{10}$$

with the boundary condition $\frac{\delta(\phi)}{|\nabla\phi|}\nabla\phi\cdot\hat{n} = \frac{\delta(\phi)}{|\nabla\phi|}\frac{\partial\phi}{\partial n} = 0$. We denote the LHS of Eq.(10) by $L_1(\phi)$. With the help of the level set function ϕ and from the Appendix, Eq.(10) can be replaced by:

$$\frac{\partial\phi}{\partial t} = -L_1(\phi) \tag{11}$$

The final solution of Eq.(11) will minimize the function $E(c_1, c_2, \phi)$ as shown in the Appendix. Because c_1 and c_2 are constants, we have $\frac{\partial E}{\partial c_1} = \frac{\partial E}{\partial c_2} = 0$. Therefore, $c_1(\phi) = \frac{\int u_0 H(\phi)dxdy}{\int H(\phi)dxdy}, c_2(\phi) = \frac{\int u_0(1-H(\phi))dxdy}{\int(1-H(\phi))dxdy}$. However in image processing, sometimes we are also interested in the texture inside each region. Therefore, instead of approximating the image in each region by a constant, we will use a linear planar surface, $u(x, y) = a + b\cdot x + c\cdot y$, to approximate the inside of each region in the next section. Here a, b, c are constants.

4 Modeling the Roof Edges

In the MS model, the second term in Eq.(5) leads u to be smooth in each region. However $|\nabla u|$ will become very large across the boundary line. Therefore the MS model can be used to detect discontinuities in the image surface. This kind of discontinues is referred to as **step edges**. They can also be detected by the Chan-Vese (CV) model due to the fact that the variation of the image intensity across the regions becomes very large if the boundaries are step edges. However there is also the case that the image is continuous but its first order derivatives are discontinuous along certain lines. That is there is a step edge in the first order derivative functional space. This kind of discontinuities is called **roof edges**. However roof edges are hard to detect by the classical MS model because it does not contain second order derivative term ($\int \nabla \cdot \nabla u \; dxdy$). In this paper, we follow the idea of [7] and parameterize the prior energy $U(u)$ in Eq.(2) as follows:

$$U(u) \propto \begin{cases} g_1(|\nabla u|^2) & \text{if} \quad i \text{ and } j \text{ belong to the same plane} \\ d(u_i, u_j) & \text{if} \quad i \text{ and } j \text{ belong to different planes} \end{cases}$$

with $d(u_i, u_j) = (a_i - a_j)^2 + (b_i - b_j)^2 + (c_i - c_j)^2$. It is clear that if $g_1(|\nabla u|^2) = |\nabla u|^2$ and ignoring the second term in $U(u)$, we will regain the classical MS model. The second term of $U(u)$ corresponds to the differences between the parameters of the two planar planes. Eq.(5) becomes:

$$E(u, C) = \int (u - u_0)^2 dxdy + \mu \int g_1(|\nabla u|^2) dxdy + \nu |C| + \int d(u) dxdy. \quad (12)$$

Where $g_1(x)$ must satisfy the following conditions [3, 9–11, 15]:(a) $g(x) = g(-x)$ (b) $g'(x) = 2xh(x)$. For the two-phase case, we can write the RHS of Eq.(12) in the form similar to Eq.(7) with

$$\begin{aligned} F = & \alpha_1(a_1 + b_1 x + c_1 y - u_0)^2 H(\phi) + \alpha_2(a_2 + b_2 x + c_2 y - u_0)^2 (1 - H(\phi)) + \\ & \mu g_1(|\nabla u|^2) H(\phi) + +\mu g_1(|\nabla u|^2)(1 - H(\phi)) + +(\nu + d(u_1, u_2))\delta(\phi)|\nabla \phi|. \quad (13) \end{aligned}$$

Substituting Eq.(13) into Eq.(9) we obtain following Euler-Lagrange equation:

$$\begin{aligned} & \delta(\phi)[-(\nu + \mu d(u_1, u_2))\nabla \cdot \frac{\nabla \phi}{|\nabla \phi|} + \alpha_1(a_1 + b_1 x + c_1 y - u_0)^2 + \mu g_1(|\nabla u_1|^2) \\ & -\alpha_2(a_2 + b_2 x + c_2 y - u_0)^2 - \mu g_1(|\nabla u_2|^2) = 0 \quad (14) \end{aligned}$$

with the same boundary condition as Eq.(10). We denote the LHS of Eq.(14) by $L_2(\phi)$.

Following the Appendix, the time dependent evolution equation of the level set $\phi(x, y, t)$ can be expressed as $\frac{\partial \phi}{\partial t} = -L_2(\phi)$. A minimal value of E can be obtained by using this level set gradient technique.

We can calculate a_i, b_i, c_i $(i = 1, 2)$ via the following equations ($\alpha_1 = \alpha_2 = 1, i \neq j$ and $i, j \in \{1, 2\}$)

$$\begin{aligned} & a_i \int (H(\phi) + |\nabla H(\phi)|) dxdy + b_i \int H(\phi) x dxdy \\ & + c_i \int y H(\phi) dxdy = \int (u_0 H(\phi) + a_j |\nabla H(\phi)|) dxdy \\ & a_i \int x H(\phi) dxdy + b_i \int [(x^2 + 2\mu h(|\nabla u_1|)) H(\phi) + |\nabla H(\phi)|] dxdy \end{aligned}$$

$$+c_i \int xyH(\phi)dxdy = \int [xu_0H(\phi) + b_j|\nabla H(\phi)|]dxdy \tag{15}$$

$$a_i \int yH(\phi)dxdy + +b_i \int xyH(\phi)dxdy + c_i \int [(y^2 + 2\mu h(|\nabla u_1|))H(\phi)$$

$$+|\nabla H(\phi)|]dxdy = \int (yu_0H(\phi) + c_j|\nabla H(\phi)|)dxdy$$

5 Experimental Results

We have implemented the above model using one level set function. In Fig.1, the segmentation results of three Chinese characters are shown. It is clear that our method gives good segmentation results. In Fig.2, we compare the segmentation results by using the CV model and our model. Again it is clear that our model gives a much better segmentation results than the CV model. In Fig.(3), more examples of segmentation are shown. In the first row of Fig.3, the CV model cannot produce a good segmentation for the lower part of the image. However our model can provide much better results. In the second row of Fig.3, the CV model mis-identified some background with foreground (bone). This is due to the fact that the CV model considers two regions as the same region if the differences of pixel intensities of two regions are not large. However in our model, we also consider the first derivative which can distinguish some small differences that cannot be detected by the CV model.

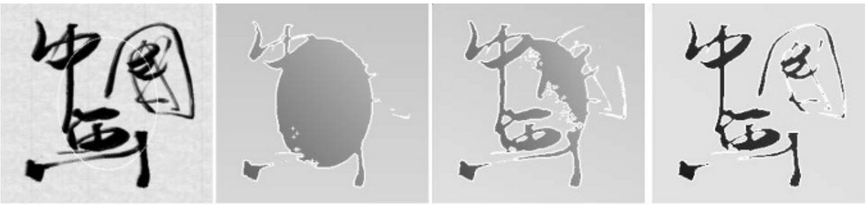


Fig. 1. The images show the evolution of the $\phi(x, y, t)$. In the calculation, we have approximated the two phases by two planar surfaces.

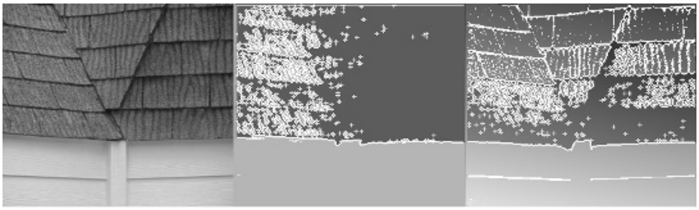


Fig. 2. From left to right: The first one is the original image, the middle is the segmentation results from the CV model, the last one is the results of our model. In the calculation, $\nu = 0.0001\sigma^2$ and σ^2 is the variance of the input image.

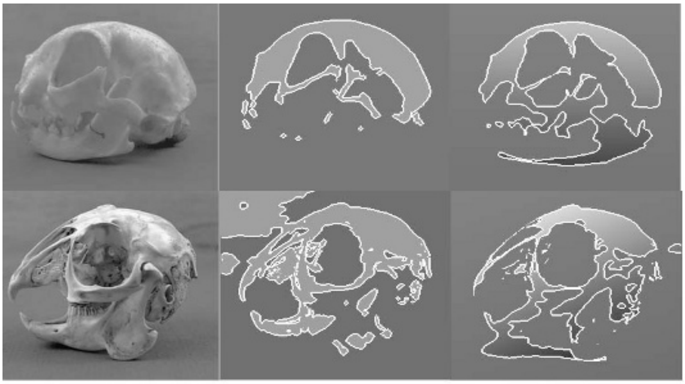


Fig. 3. From left to right: The first one is the original image, the middle is the segmentation results from the CV model, the last one is the results of our model. In the calculation, $\nu = 0.5\sigma^2$ and σ^2 is the variance of the input image.

6 ν-Dependence in the Chan-Vese Model

It has been observed in [2, 4] that the segmentation results depend on the ν value in the CV model. It has been pointed out in [4] that when ν becomes very large, the segmentation result of the Chan-Vese model is bad. This can be easily understood from the CV model (see Eq.(7)). For a fixed boundary, if ν is very large, then the length will contribute largely to the total energy $E(u, C)$. On the other hand, if ν is very small, then the length will contribute very little to the total energy. Therefore, if ν is very large, we cannot expect a very long boundary; otherwise $E(u, C)$ will be large. On the other hand, if ν is very small, then we could find a long boundary. Thus, for large objects, we have to choose a very small ν.

We notice that the dependence on ν in our model is not as strong as in the CV model. This is due to the fact that in Eq.(13), the length term contains both ν and $d(u_1, u_2)$. As long as $d(u_1, u_2)$ is bigger than ν, our segmentation results will depend not so much on the value of ν. We can understand this in the following way: if d is very small, that is the differences between the parameters of two regions are small, thus these two regions tend to be one region. Therefore the length term will be short. On the other hand, if d is large, the two regions are separated, accordingly the length term will be long.

In Fig.4, the segmentation results for both the CV and our models are shown for different ν values. It is clear that as ν becomes small, our model depends very little on the values of ν. However the CV model depends significantly on the ν value. This is another advantage of our model. In the following we will establish a limit on the value of ν. Suppose the boundary is empty, $|C| = 0$, then the total MS model energy of the image (see Eq.(5)) is $\mathbf{N}\sigma^{\mathbf{2}}$. Here σ^2 is the variance of the input image and N is the total number of pixels. Because the first two terms of Eq.(5) are not less than zero, the energy due to the boundary $|C|$ must be

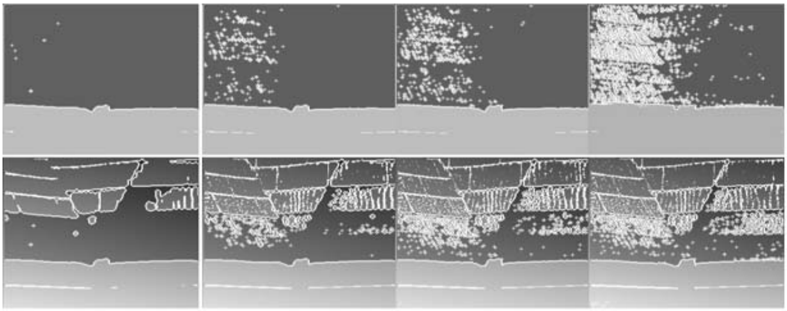


Fig. 4. From left to right: $\nu = 0.5, 0.1, 0.01, 0.001(\sigma^2)$ and σ^2 is the variance of the input image. The first row is the results of CV model and the second row is our results. The original image is in Fig. 2.

less than the total energy $\mathbf{N}\sigma^{\mathbf{2}}$, i.e. $\nu|\mathbf{C}| \leq \mathbf{N}\sigma^{\mathbf{2}}$. As the maximum value of the total length $|C|$ is of order N. Therefore we have $\nu \leq \sigma^2$.

7 Conclusions

We have connected the MS model to the Markov random field theory. It is shown that these two models are equivalent. In our model, we have approximated the bounded regions by piecewise planar functions which can preserve certain degree of texture in each region. The classical MS model cannot detect the roof edges because it does not include second order derivative terms. To overcome this problem, we have generalized the MS model to include differences of parameters of the planar planes. It is found that this new model can detect roof edges which is hard to detect by other models. We have also shown that this new model is less dependent on the parameter ν (scale factor) than the original MS model. This is important for image segmentation.

We have applied this model to many images and the segmentation results are encouraging. We have also proved that the gradient projection method will provide us a minimal energy of the input image. This proof is very general and it does not depend on the model.

References

1. G. Aubert and P. Kornprobst, Mathematical Problems in Image Processing: Partial differential equations and the Calculus of Variations, Vol. 147 of Applied Mathematical Sciences, Springer-Verlag, 2002.
2. T. F. Chan and L. A. Vese, Active Contours without edges, IEEE transactions on Image Processing, 2001, 10(2): 266-277.
3. P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, Deterministic edge-preserving regularization in computer imaging, IEEE Trans. Image Process. 6, 1997, 298-311.

4. S. Gao and Tien D. Bui, Image Segmentation and Selective Smoothing by Using Mumford-Shah Model, to be published in IEEE Transaction on Image Processing.
5. S. Gao and Tien D. Bui, A new image segmentation and smoothing model, Proc. of IEEE int. Symposium on Biomedical Imaging: From Nano to Macro, pp. 137-140, Arlington, V.A., April 15-18, 2004.
6. B. R. Lee, A. Ben Hamza and H. Krim, An active contour model for image segmentation: a variational perspective. Proc. IEEE international conference on acoustics speech and Signal processing, Orlando, May 2002.
7. S. Z. Li, Roof-Edge Preserving Image Smoothing Based on MRFs, IEEE Transactions on Image Processing, Vol. 9, No. 6, June 2000, pp.1134-1138.
8. D. Mumford, and J. Shah, Optimal approximation by piecewise smooth functions and associated variational problems. Comm. Pure Appl. Math. 42 (1989) 577 -685.
9. F.A. Pellegrino, W. Vanzella, and V. Torre, Edge Detection Revisited, IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics, Vol. 34, No. 3, June 2004, pp.1500-1517.
10. J. A. Sethian, Level Set Methods and Fast Marching Methods, Cambridge University Press, 1999.
11. S. Teboul, L. Blanc-Feraud, G. Aubert, and M. Barlaud, Variational Approach for Edge-Preserving Regularization Using Coupled PDEs, IEEE Transactions on Image Processing, Vol. 7, No. 3, March 1998 387-397.
12. A. Tsai, A. Yezzi, and Alan S. Willsky, Curve Evolution Implementation of the Mumford-Shah Functional for Image Segmentation, Denoising, Interpolation, and Magnification, IEEE Tran. on Image Processing, Vol. 10 (8), 2001, 1169-1186.
13. L. Vese and T. F. Chan, A Multiphase Level Set Framework for Image Segmentation Using the Mumford and Shah Model, International Journal of Computer Vision 50(3), 271-293, 2002.
14. L. A. Vese, Multiphase Object Detection and Image Segmentation, in "Geometric Level Set Methods in Imaging, Vision and Graphics", S. Osher and N. Paragios (eds), Springer Verlag, 2003, pp. 175-194.
15. R.A. Weisenseel, W.C. Karl, D.A. Castanon, A region-based alternative for edge-preserving smoothing, Proceedings of the International Conference on Image Processing, 2000. pp. 778-781. Vancouver, BC, Canada.
16. S. Geman, and D. Geman, Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images. IEEE Trans. on PAMI. 6 (1984) 721-741.
17. A. Blake and A. Zisserman, Visual Reconstruction, The MIT Press Cambridge, Massachusetts, 1987.
18. W. Vanzella, F.A. Pellegrino, and V. Torre, Self-Adaptive Regularization, IEEE Transactions on PAMI, Vol. 26, No. 6, June 2004, pp.804-809.

Appendix

To solve Eq.(9), we can use the gradient technique. That is, we change Eq.(9) to the following time dependent equation

$$\frac{\partial \phi}{\partial t} = G(L(\phi)). \tag{16}$$

If this equation has a stable solution as time goes to infinity, that is $\frac{\partial \phi}{\partial t}|_{t \to \infty} = 0$, then we have $G(L(\phi)) = 0$. If as long as $G(L(\phi)) = 0$, $L(\phi) = 0$, then the solution of Eq.(16) is also a solution of Eq.(9). From Eq.(9), we have

$$\begin{aligned}\frac{\partial E}{\partial t} &= \int_{\Omega}[\frac{\partial F}{\partial \phi}\frac{\partial \phi}{\partial t} + \frac{\partial F}{\partial \phi_x}\frac{\partial \phi_x}{\partial t} + \frac{\partial F}{\partial \phi_y}\frac{\partial \phi_y}{\partial t}]dxdy \\ &= \int_{\Omega}[\frac{\partial F}{\partial \phi}\phi_t + \frac{\partial F}{\partial \phi_x}\frac{\partial \phi_t}{\partial x} + \frac{\partial F}{\partial \phi_y}\frac{\partial \phi_t}{\partial y}]dxdy = \int_{\Omega}[\frac{\partial F}{\partial \phi}\phi_t + \boldsymbol{A}\cdot\nabla(\phi_t)]dxdy \\ &= \int_{\Omega} L(\phi)\phi_t dxdy + \int_{\delta\Omega}(\phi_t)\boldsymbol{A}\cdot\hat{n}dl. \end{aligned} \tag{17}$$

with $\boldsymbol{A} = (\frac{\partial F}{\partial \phi_x}, \frac{\partial F}{\partial \phi_y}), \phi_t = \frac{\partial \phi}{\partial t}$. $\hat{n}$ is the normalized normal of curve $\delta\Omega$. $L(\phi)$ is the LHS of Eq.(9). Here we have used Gaussian theorem in the above derivation. The last term in the last line of Eq.(17) is zero due to the boundary conditions in Eq.(9). Using Eqs.(16) and (17) and taking $G(L(\phi)) = -L(\phi)$, we have

$$\frac{\partial E}{\partial t} = \int L(\phi)G(L(\phi)) = -\int L^2(\phi)dxdy \leq 0. \tag{18}$$

Therefore, the solution of Eq.(16) is the solution of the Eq.(8) which will minimize the energy functional $E(\phi)$ in Eq.(8).

A Dynamic Stroke Segmentation Technique for Sketched Symbol Recognition

Vincenzo Deufemia and Michele Risi

Dipartimento di Matematica e Informatica
Università di Salerno,
84084 Fisciano (SA), Italy
{deufemia,mrisi}@unisa.it

Abstract. In this paper, we address the problem of ink parsing, which tries to identify distinct symbols from a stream of pen strokes. An important task of this process is the segmentation of the users' pen strokes into salient fragments based on geometric features. This process allows users to create a sketch symbol varying the number of pen strokes, obtaining a more natural drawing environment. The proposed sketch recognition technique is an extension of LR parsing techniques, and includes ink segmentation and context disambiguation. During the parsing process, the strokes are incrementally segmented by using a dynamic programming algorithm. The segmentation process is based on templates specified in the productions of the grammar specification from which the parser is automatically constructed.

1 Introduction

Sketches greatly simplify conceptual design activities through abstract models that let designers express their creativeness, and focus on critical issues rather than on intricate details [9]. Due to their minimalist nature, i.e., representing only what is necessary, they enhance collaboration and communication efficiency.

Underlying a sketch-based user interface several processes can be activated. These include the processing of pen strokes, recognition of symbols, stroke beautification, reasoning about shapes, and high-level interpretation. The sketch understanding tasks are not trivial because recognizing the meaningful patterns implied by a user's pen stroke must be flexible enough to allow some tolerance in sketch recognition, but sufficiently constrained not to accept incorrect patterns. Furthermore, the context in which a particular stroke or group of strokes appears considerably influences the interpretation of that stroke. From a visual language point of view, this means that the interpretation of a graphical object is strongly influenced by the objects surrounding it. Moreover, semantically different objects might be graphically represented by identical or apparently similar symbols.

Another important issue in sketch understanding concerns ink parsing, which refers to the task of grouping and segmenting the user's strokes into clusters of intended symbols. This allows users to create a sketch symbol varying the number of pen strokes, obtaining a more natural drawing environment.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 328–335, 2005.

Several systems for sketch recognition constrain users to draw an entire symbol as single stroke [8,9], or to draw only strokes representing single primitive shapes such as lines, arcs, or curves [7,14]. In other systems prior to parsing, the input sketch is segmented into line and arc segments allowing symbols to be drawn with multiple pen strokes, and a single pen stroke to contain multiple symbols [2,11]. Different approaches to segmentation have been proposed, some systems segment the strokes by using their curvature and speed information [2,11], Saund uses both local features (such as intersections and curvatures) and global features (such as closed paths) to locate breakpoints of a stroke [12], whereas Yu applied the mean shift procedure to approximate strokes [13]. Hse *et al.* [6] have presented an optimal segmentation approach based on template that does not suffer of over- and under-segmentation of strokes. In particular, given a sketched symbol *S* and a template *T*, the algorithm finds a set of breakpoints in *S* such that the fitting performed according to *T* yields the minimum fit error. The templates *T* can be of two types, one specifies a sequence of lines and ellipses and the other specifies the number of lines and ellipses.

Segmentation is a basic problem that has many applications for digital ink capture and manipulation, as well as higher-level symbolic and structural analyses. As an example, the structural information generated by the segmentation process can be useful for the beautification of the symbols [7], for developing a user interface with which to interact with sketched ink.

In this paper, we present a sketch recognition technique that takes into account both the problem of context based ambiguity resolution and ink segmentation. The sketch parser relies on an extension of LR parsing techniques and is automatically generated from a grammar specification. The proposed segmentation technique dynamically segments the strokes during the parsing process by using an extended version of the optimal pen strokes segmentation technique proposed by Hse *et al.* [6]. The template given in input to the algorithm is a sequence of primitive shapes iteratively extracted from grammar productions. Thus, the parser's context drives the segmentation process of the strokes.

The paper is organized as follows. We first discuss the formalism for describing sketch languages, and the parsing approach underlying the proposed framework. Successively, we present a dynamic segmentation technique that integrated into the parsing algorithm solves the problem of multi-stroke recognition. In section 4 we describe an example of application of the parsing algorithm on hand-drawn circuit diagrams. Finally, the conclusion and further research are discussed in Section 5.

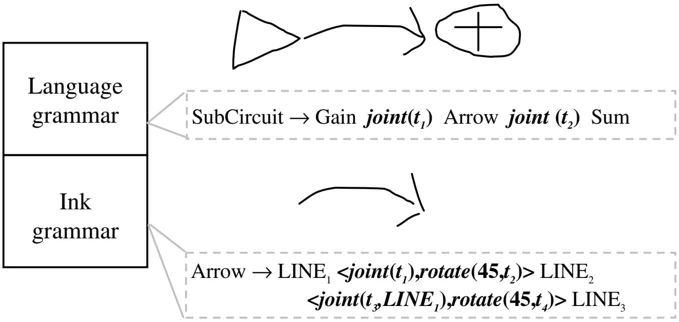
2 A Grammar-Based Sketch Parsing Approach

Because we use an extension of LR parsing technique [1], we describe sketch languages using the formalism of *eXtended Positional Grammars* (XPG, for short) [4]. XPGs represent a direct extension of context-free string grammars, where more general relations other than concatenation are allowed. A sentence is conceived as a set of symbols with attributes. Such attributes are also determined by the relationships holding among the symbols. Thus, a sentence is specified by combining symbols with relations. In particular, the productions have the following format:

$$A \rightarrow x_1 \boldsymbol{R_1} x_2 \boldsymbol{R_2} \ldots x_{m-1} \boldsymbol{R_{m-1}} x_m$$

where each $\boldsymbol{R_j}$ define a sequence of relation between x_{j+1} and x_{j-i}, with 1≤i<j, by means of a threshold t_j.

An XPG for modeling a sketch language L can be logically partitioned into two XPG grammars. The first, named *ink grammar*, defines the symbols of the language L as geometric compositions of primitive objects, i.e., patterns that cannot be recognized as a combination of other objects and must be recognized directly, such as line segments and elliptical arcs. For example, a production specifying an *Arrow* symbol is shown in the bottom of Figure 1 together with a sketch matching such production. The second grammar, named *language grammar*, specifies the sentences of the language as compositions of shapes defined in the ink grammar through spatial relations. The production at the top of Figure 1 defines a *SubCircuit* object as the composition of three language symbols: *Gain*, *Arrow* and *Sum*, and shows a sketch matching the production.



SubCircuit → Gain ***joint***(t_1) Arrow ***joint*** (t_2) Sum

Arrow → $LINE_1$ <***joint***(t_1),***rotate***(45,t_2)> $LINE_2$
<***joint***(t_3,$LINE_1$),***rotate***(45,t_4)> $LINE_3$

Fig. 1. The two levels of XPG grammar specification.

The recognition of line segments is performed with the least square fitting [5], whereas the elliptical arc fitting is performed with the technique proposed in [10].

The definition of XPGs has been strongly influenced by the need of having an efficient parser able to process the generated languages. Due to their analogy with string grammars, it has been natural the use of LR parsing techniques [4]. The result was a parser that scans the input in a non-sequential way, driven by the relations used in the grammar. In order to guide the scanning of the input symbols, a new column next is added to the usual action and goto parts of an LR parsing table. For each state of the parser, this column contains an entry with information to access the next symbol to be parsed. This information is derived by the relations of the grammar productions, during the parser generation. Thus, during the parsing process, a sketch parser generates a sequence of calls to a function *Fetch_Stroke* that linearizes the input at run-time [3].

Figure 2 shows the structure of the recognition process. During the parsing of a sketch diagram, a rank value is computed by combining the accuracy of the strokes forming the sketch and of their spatial relations. Thus, the output of the parser is a *probabilistic parse forest* where each tree in corresponds to an interpretation of the sketch sentence. Each node of a tree has associated a probability representing the rank of the stroke interpretation associated to the leaves of its subtree. Such trees can be analyzed to obtain a rank of the interpretations by considering the probability associated to the roots of the trees and the number of language symbols recognized, similarly to what done for natural languages.

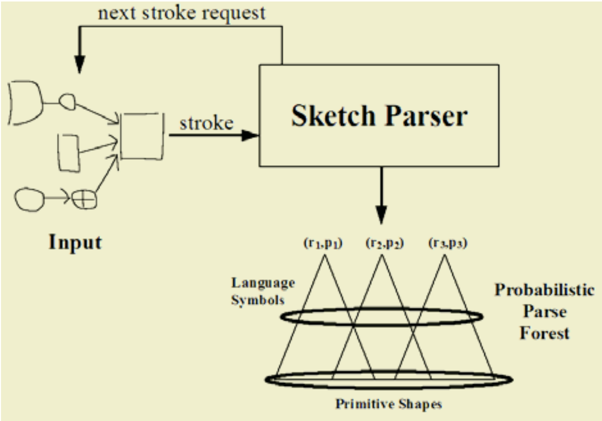


Fig. 2. The sketch parsing approach.

3 A Dynamic Ink Segmentation Technique

The previous sketch parsing approach does not allow pen strokes to represent any number of shape primitives connected together, since the function *Fetch_Stroke* finds in the input sketch a single stroke that matches the primitive shape specified in a grammar production.

This problem can be overcome by segmenting the strokes when the parser cannot proceed in the recognition of the sketch. Thus, the sequences of primitive shapes specified in the grammar productions, together with their geometric relationships, guides the segmentation algorithm to identify the points for dividing the strokes into different primitives.

More formally, given a stroke *S* formed by a set of *m* data points, a template *p* formed by a sequence of primitive shapes, and an array *r* of relations between such shapes, the segmentation algorithm finds a subset of the *m* points that matches with the pattern *p* and the array *r* yielding the minimum fit error.

This “fitting to a template” problem can be optimally solved by using the dynamic programming approach proposed by Hse *et al.* [6], and considering both the similarity between segments and patterns, and the quality of the relations between the segments. Thus, the output of the algorithm is the best segmentation according to the best fit shape error and the best accuracy of the shape relations.

Let *d(m,k,p,r)* be the minimum error segmentation, where *m* is the number of data points describing the stroke S, *k* is the number of breakpoints to be determined by the segmentation process (the start value is *k= p.len-1*), *p* is the template of primitive shapes, and *r* is the array of relations specified in the XPG productions between the shapes in *p*.

The recursive definition for segmentation of *S* is given in the following.

$$d(m,k,p,r)=\begin{cases} f(S,0,m,p[1],r[1]) & \text{if } k=0 \\ \min\limits_{k<i<m}\{d(i,k-1,p[1..p.len],r[1..r.len])+f(S,i,m,p[p.len],r[r.len])\} & \text{if } k>0 \end{cases}$$

f(S,i,m,p[j],r[n]) is the *segmentation error function* defined as:

$$f(S,i,m,p,rel)=\frac{fitting(S,i,m,p)}{relation(rel,p,rel.th)}$$

This function is calculated considering the fitting of a segment from *i-th* point to *m-th* point in S using *p*, and considering the accuracy of the relation *rel* between the current matched symbol *p* and the previous one by means of the threshold *rel.th*.

When the *Fetch_Stroke* function fails to find a stroke accurately related with a previously parsed stoke, the last visited stroke *s* could contain multiple symbols. Thus, the segmentation process is activated on *s* in order to calculate the breakpoint that divide *s* into two strokes s_1 and s_2, and such that they have a good compromise between their fitting shape error and the accuracy in the relations involving them. Successively, the parser considers s_1 as the last visited stroke, and the *Fetch_Stroke* function returns s_2 as the next stroke to be parsed. The segmentation process on the stroke *s* is iterated until both *m>p.len* and *Fetch_Stroke* fails to find a stroke not in *s* and accurately related with a previously parsed stoke.

For example, a square can be drawn as a single pen stroke, or as two separate strokes, or even as three or four strokes (Fig. 3(a)-(d), respectively).

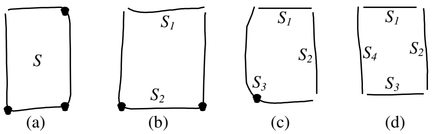


Fig. 3. Segmentation of a square.

The production used to recognize the square is:

Square → LINE$_1$ ***<joint(t$_1$), rotate(90, t$_2$)>*** LINE$_2$ ***<joint(t$_1$), rotate(90, t$_2$)>***
LINE$_3$ ***<joint(t$_1$), rotate(90, t$_2$), joint(t$_1$, LINE$_1$)>*** LINE$_4$

If the square is drawn using four strokes then the production is correctly reduced without segmentation since the strokes returned by *Fetch_Stroke* match the relations in the production.

During the parsing of the square in Figure 3(c), the *Fetch_Stroke* matches the first three lines of the production with the strokes S_1, S_2 and S_3, but it fails to find the symbol LINE$_4$. Thus, a segmentation process is activated on the last visited stroke S_3. In particular, the algorithm calculates the value $d(m,1,[L,L],[\{joint(t_1),rotate(90,t_2)\},\{joint(t_1), rotate\ (90,t_2), joint(t_1)\}])$ determining the breakpoint shown as a filled circle in Figure 3(c). The parser continues the recognition of the square by matching the segmented stroke S_3 with LINE$_3$ and LINE$_4$. Similarly, the parser recognizes the squares in Figures 3(a) and 3(b) by segmenting the strokes S and S_2 with $d(m,3,[\mathrm{L,L,L,L}],[\varnothing,\{\mathit{joint}(t_1), \mathit{rotate}(90,t_2)\}, \{\mathit{joint}(t_1), \mathit{rotate}(90,t_2)\},\{\mathit{joint}(t_1), \mathit{rotate}(90,t_2), \mathit{joint}(t_1)\}])$ and $d(m,2,[\mathrm{L,L,L}],[\{\mathit{joint}(t_1), \mathit{rotate}(90,t_2)\},\{\mathit{joint}(t_1),\mathit{rotate}(90,t_2)\},\{\mathit{joint}(t_1), \mathit{rotate}(90,t_2), \mathit{joint}(t_1)\}])$, respectively.

It worth noting that the time and space complexity of the segmentation process is the same of the Hse's algorithm [6] since when the parser segments a stroke with the pattern $p=[p_1,\ldots,p_{n-1},p_n]$, the segmentation value of the pattern $p_r=[p_1,\ldots,p_{n-1}]$ has already been calculated previously.

4 Segmenting Hand-Drawn Electrical Circuits

In this section we show how the proposed dynamic stroke segmentation technique works during the recognition of hand-drawn circuit diagrams.

The symbols in the circuit domain are given in Figure 4.

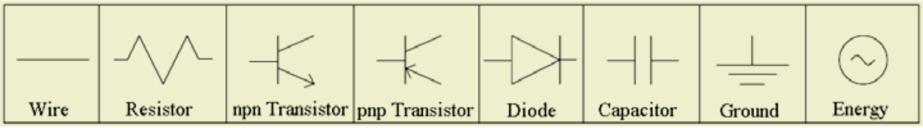


Fig. 4. The visual symbols in the circuit language.

In the following we describe some productions of the ink grammar for modeling the resistor, the wire, the capacitor and the ground symbols. In particular, the resistor symbol starts and ends with a line, and in the middle it is composed of a sequence of at least four oblique lines forming a wave.

Resistor → LINE ***<joint(t_1), rotate(225, t_2)>*** Wave ***<joint(t_1), rotate(135, t_2)>*** LINE
Wave → LINE$_1$ ***<joint(t_1), rotate(45, t_2)>***
LINE$_2$ ***<joint(t_1), rotate(135, t_2), >***
LINE$_3$ ***<joint(t_1), rotate(45, t_2), parallel(LINE$_{1,}$ t_3)>***
LINE$_4$ ***<joint(t_1), rotate(135, t_2), parallel(LINE$_2$, t_3)>*** Multiwave
Multiwave → LINE ***<joint(t_1), rotate(45, t_2), parallel(LINE(-1), t_3)>*** Multiwave
Multiwave → LINE

Wire → Wire ***<intersect(t_1)>*** LINE
Wire → LINE

Capacitor → LINE$_1$ ***<perpendicular(t_1), intersection(t_2)>*** LINE$_2$ ***<parallel(t_3), length(1.0, t_4)>***
LINE$_3$ ***<perpendicular(t_1), intersection(t_2)>*** LINE$_2$

Ground → LINE ***<perpendicular(t_1), intersection(t_2)>*** Multiline
MultiLine → LINE ***<parallel(t_3),length(0.7, t_4), centered(t_5)>***
LINE ***<parallel(t_3),length(0.7, t_4), centered(t_5)>*** Subline
Subline → LINE ***<parallel(t_3),length(0.7, t_4), centered(t_5)>*** Subline
Subline → LINE

Notice that the LINE symbol may further be refined in order to also consider multi-stroke segments.

The following productions represent some of the language grammar productions for the circuit language.

Circuit → SubBlock
SubBlock → SubBlock <***joint***(t_1)> Wire <***joint***(t_1)> Component
SubBlock → SubBlock <***joint***(t_1)> Wire
SubBlock → SubBlock <***any***> Component
SubBlock → Component
Component → Resistor
→ Capacitor
→ Ground
→ Diode
→ pnpTransistor
→ npnTransistor
→ Energy

Fig. 5 shows an electric circuit and the segmentation of a resistor symbol. In particular, when the parser starts the recognition of the resistor it looks for a line segment. Since the resistor symbol has been drawn with two complex strokes, the line fitting algorithm fails on matching the first stroke with a line. Thus, the parser segments the stroke identifying a breakpoint that divides the stroke into two lines according to the grammar productions. The optimal breakpoint is point 1 and the parser proceeds the recognition on the remaining part of the segmented stroke. In particular, driven by the production defining the symbol *Wave* the parser segments the stroke into four line segments identifying the breakpoints 2, 3 and 4. Successively, the parser segments the second stroke forming the resistor by considering the production defining the symbol *Multiwave*.

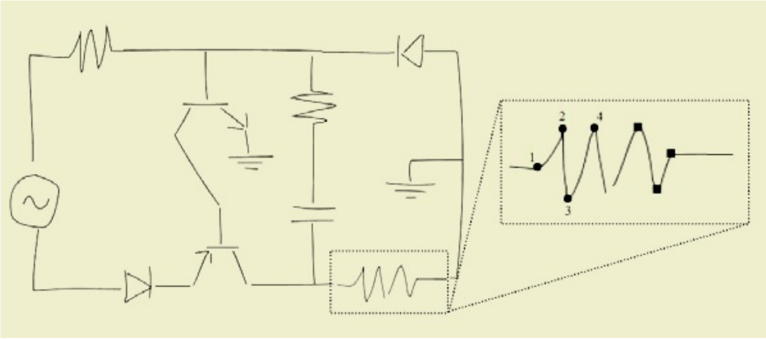


Fig. 5. An electric circuit with a resistor drawn with two strokes and segmented by the proposed algorithm during its parsing.

5 Conclusion

We have presented a sketch parsing approach that includes a context based ink segmentation technique. Indeed, template derived from grammar productions drives the segmentation process. The approach is designed to enable natural sketch-based computer interaction since it allows for multiple symbols to be drawn in the same stroke, and allows individual symbols to be drawn in multiple strokes.

We have integrated the proposed approach in the SketchBench system [3], a tool supporting the early phases of sketch language modeling, such as shape modeling and grammar specification, and the generation of the final parser.

We are currently conducting an evaluation test of the segmentation technique.

References

1. Aho, A., Sethi, R., and Ullman, J.: Compilers Principles ,Techniques, and Tools. Addison-Wesley Series in Computer Science. 1987.
2. Calhoun, C., Stahovich, T.F., Kurtoglu, T. and Kara, L.B.: Recognizing Multi-Stroke Symbols, in AAAI Symposium - Sketch Understanding, (2002), 15–23.
3. Costagliola, G., Deufemia, V., Polese, G., and Risi, M.: A Parsing Technique for Sketch Recognition Systems, in Proceedings of IEEE Symposium VL/HCC'04, (Rome, September 2004), IEEE Press, 19–26.
4. Costagliola, G. and Polese, G.: Extended Positional Grammars, in Proceedings of IEEE Symposium VL'00, (Seattle, WA, September 2000), IEEE Press, 103–110.
5. Duda, R.O. and Hart, P.E.: Pattern Classification and Scene Analysis. Wiley Press, New York, 1973.
6. Hse, H., Shilman, M., and Newton, A.R.: Robust Sketched Symbol Fragmentation using Templates, in Proceedings of IUI'04 (Madeira, Portugal, January 2004), ACM Press, 156–160.
7. Igarashi, T., Matsuoka, S., Kawachiya, S. and Tanaka, H.: Interactive beautification: A technique for rapid geometric design, in Proceedings of UIST'97, 105–114.
8. Kimura, T.D., Apte, A., and Sengupta, S.: A graphic diagram editor for pen computers. Software Concepts and Tools, 1994, 82–95.
9. Landay, J. and Myers, B.: Sketching interfaces: Toward more human interface design. IEEE Computer, 34(3), 2001, 56–64.
10. Pilu, M., Fitzgibbon, A. and Fisher, R.: Direct Least-Square Fitting of Ellipses. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21 (5), 1999, 476–480.
11. Sezgin, T.M., Stahovich, T. and Davis, R.: Sketch Based Interfaces: Early Processing for Sketch Understanding, in Procs of PUI'01, (Orlando, 2001).
12. Saund, E.: Finding Perceptually Closed Paths in Sketches and Drawings. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25 (4), 2003, 475–491.
13. Yu, B.: Recognition of Freehand Sketches using Mean Shift, in Proceedings of IUI'03, (Miami FL, 2003), 204–210.
14. Weisman, L.: A foundation for intelligent multimodal drawing and sketching programs, Master's thesis, MIT, 1999.

Application of Wavelet Transforms and Bayes Classifier to Segmentation of Ultrasound Images

Paweł Kieś

Institute of Fundamental Technological Research, Polish Academy of Sciences,
21 Świętokrzyska Str., 00-049 Warsaw, Poland
pkies@ippt.gov.pl

Abstract. An approach for segmentation of ultrasound images using features extracted by orthogonal wavelet transforms that can be used in an interactive system is proposed. These features are the training data for the K-means clustering algorithm and the Bayes classifier. The result of classification is improved by using neighbourhood information.

1 Introduction

The use of ultrasound scanners in medical diagnostics becomes more and more popular. The equipment develops and becomes cheaper and more accessible for doctors. It more and more often is used in applications usually reserved for an expensive equipment like computer tomography or magnetic resonance devices. An example of this kind of application is an automatic segmentation of images used in 3D visualisation systems [1].

The problem with ultrasound images is the fact that the most usable information is contained not in gray levels of an image, but in pixel patterns called textures [4], defined by the gray level distribution in a neighbourhood of the investigated pixel. Moreover, the usable information can be stored at various levels of resolution of the texture. A very efficient method for multiresolution features extraction used for a texture segmentation is a discrete wavelet transform [2], that one divides into two types: orthogonal and nonorthogonal [7, 8, 11].

In Section 2 we describe a typical application framework where the proposed approach can be applied. In section 3 a method of a Region Of Interest (ROI) reduction is proposed. Next, in Section 4 we describe an orthogonal wavelet transform as a source of features and the feature extraction method in Section 5. The classification method is proposed in Section 6. The experimental results are presented in Section 7 and conclusions in Section 8.

2 Application Description

We assume that the ROI (see Fig. 2, 2nd row) is a graylevel image defined by an user (a doctor) using any drawing program (e.g. GIMP or Photoshop) that provides a pen-like tool and makes it possible to draw on a layer over an ultrasound image.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 336–342, 2005.

We select a special graylevel (e.g. black) that means the omitted part of the image. The remaining graylevels mean classes of objects. Thus, the ROI can provide both kinds of information: the part of the input image to be processed and the classification of pixels used for training the classifier and testing the correctness of the classification.

We expect that the user marks by related graylevels all pixels belonging each class, especially the object and the non-object (background) for two classes (see Fig. 2.a & c). Eventually, the user can mark the pixels surrounding the border between classes (see Fig. 2.b).

We can divide the training algorithm into the following steps described in following sections:

1. ROI reduction,
2. features extraction,
3. training the classifier.

In order to classify unknown pixels at a similar ultrasound image the user should mark interesting pixels by making another ROI image. As above, he can mark the pixels surrounding the expected border between classes, maybe basing on the classification of the previous slice in a 3D imaging system. Similarly, we can divide the classification algorithm into the following steps:

1. features extraction,
2. classification,
3. refinement of the classification.

We assume a high correlation of the class assignment between neighbouring pixels. So we can refine the results by applying a Gaussian filter to the membership image for each class.

3 ROI Reduction

We suppose that the number of pixels selected for training in the way described in the previous section is highly excessive and we can try to reduct the data in order to speed-up the training algorithm and to improve the system's interactivity. The proposed method makes it possible to control the number of training examples and gives the opportunity to adjust it according to any particular costs of misclassification in each class.

The proposed probabilistic reduction algorithm is controlled by the following parameters:

- M_i – the number of pixels belonging to the ith class;
- M_R – an expected number of training pixels – can be assessed from the time response restrictions;
- η_i – the relative cost of an incorrect classification of a pixel to the ith class.

Then we can calculate the probability that a pixel from the ith class belongs to the training set

$$p_i = \frac{M_R}{M_i C} \cdot \frac{\eta_i}{\sum_{i=1}^{C} M_i \eta_i}, \tag{1}$$

and we use it to sample randomly the ROI provided by the user (see Fig. 2, 2nd row). As a result we receive a reduced ROI being used as the training set (see Fig. 2, 3rd row).

Remaining pixels of each class (not qualified to the training set) are used as the testing set, thus the ratio of testing and training pixels for the ith class is

$$\kappa_i = (1 - p_i)/p_i. \tag{2}$$

4 Wavelet Transform

In this paper we apply orthogonal wavelet transforms only, because they are simpler and, probably, more adequate for an use in interactive systems. We will compare two kinds of them: a Discrete Wavelet Transform (DWT) and a Discrete Wavelet Packet Transform (DWPT).

A simple transform consists of a low-pass (L) and high-pass (H) mirror filters applied to the image twice: by rows and by columns. As a result, the image is decomposed into 4 subband channels: LL, LH, HL, HH. The LL channel is a coarse or low-pass one. The remaining channels are detail or high-pass ones.

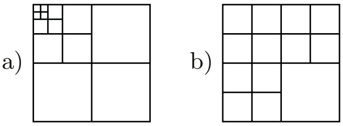


Fig. 1. Image's divisions into 13 subband channels: (a) DWT for $L = 4$, and (b) DWPT for $L = 2$.

The DWT consists of a certain number of decompositions of the coarse channel (LL) only (Fig. 1a). Its use is motivated by the fact that usually the majority of image energy goes to this channel. The DWPT decomposes possibly all channels (Fig. 1b). The decision on a particular decomposition depends on an energy measurement. A channel is decomposed if it contains a considerable part of total energy.

The DWT gives $1 + 3L$ features and the DWPT gives the number of features in the range $\langle 1{+}3L, 4^L \rangle$, where L is the level of transformation, i.e. the maximum number of applications of a simple transform to any part of the image.

5 Features Extraction

The features extraction using an orthogonal wavelet transform is composed of the following steps [9]:

1. Decompose the input image using a wavelet transform. As a result we receive an image of the same dimensions as the input image, but divided into subband channels.

2. Calculate an absolute value for all high-pass channels.
3. Apply a Gaussian filter as an envelope detection algorithm to each high-pass channel of the transform. The slope of the filter should depend on the mean size of a texture element.
4. For each pixel in the ROI determine feature values from related channels, do necessary interpolations and store as a feature matrix.
5. Standardize values of each feature in the feature matrix.

The training stage uses a feature matrix extracted from an image with correctly classified pixels, as an training data, and gives parameters of the classification method as the output.

The classification stage takes parameters of the classification method, a feature matrix extracted from an image with an unknown pixel classification and the ROI definition as the input. It gives an estimated label for each pixel in the ROI as the output.

6 Classification Method

Tissues and organs at ultrasound images usually are inhomogeneous regions. Each object of interest can be composed of textures of various visual properties. So we can use only such a classification method that can assign many types of tissues to a single class. On the other side we cannot use methods of a high numerical complexity (e.g. neural networks [6] or support vector machines [12]).

These conditions are fulfilled by an algorithm composed of the K-means clustering algorithm [5] and the Bayes classifier [3, 11]. The K-means algorithm is an iterative procedure composed of the following steps:

1. Assign all items randomly to K clusters.
2. Calculate a mean vector for each cluster.
3. Assign all items to clusters using a minimum distance method.
4. If the present class assignment differs from the previous one go to step 2.

The algorithm is run during the training stage separately for each class of correctly classified items, that gives totally KC clusters, where C is the number of classes, and K is the number of clusters in each class. It is an adjustable parameter of the algorithm.

The Bayes classifier assumes that probability distribution $p(x|\omega_j)$ for each cluster ω_j is a N-dimensional normal distribution defined by two parameters:

- the mean vector $\mu_j = \mathbf{E}\left[\mathbf{x}|\omega_j\right]$ and
- the covariance matrix $\mathcal{M}_j = \mathbf{E}\left[(\mathbf{x} - \omega_j)(\mathbf{x} - \omega_j)^T|\omega_j\right]$.

These parameters are calculated for each cluster at the end of the training stage.

In the classification stage each pixel is assigned to such a cluster that gives the maximum posterior probability, i.e. that maximizes the discriminant function:

$$g_j(x) = -1/2(x - \mu_j)^T \mathcal{M}_j^{-1}(x - \mu_j) - 1/2 \log|\mathcal{M}_j| + \log p(\omega_j), \qquad (3)$$

where $p(\omega_j) = |\omega_j|/M_i$ is the probability of the cluster ω_j, and M_i is the number of all pixels in the ith class.

7 Experimental Results

For our experiments we chose three types of ultrasound images (Fig. 2, 1st row), that could appear in the application described in Section 2.

We applied two transforms: DWT and DWPT, based on the 2nd order Daubechies wavelet [2] with subband channels at Fig. 1, as the source of features.

The number of clusters in each class in the K-means algorithm was $K = 20$. The total number of pixels in the reduced ROI (the training set) was $M_R = 8000$, equally divided into both classes ($\eta_1 = \eta_2 = 1$). The structure of the training and testing sets is shown in Table 1.

Table 1. Structure of the training and testing sets. For the ith class are given: M_i – number of all pixels and κ_i – ratio of the power of testing and training set (see Section 3).

Image	Class #1		Class #2	
at Fig. 2	M_1	κ_1	M_2	κ_2
(a)	58738	13.7	64326	15.1
(b)	34323	7.6	20657	4.2
(c)	70650	16.7	10372	1.6

Table 2. Results of experiments.

Image at Fig. 2	Transform	$[P_1, P_2]$	$[P_{1r}, P_{2r}]$
(a)	DWT	$[0.982, 0.989]$	$[0.985, 0.992]$
	DWPT	$[0.988, 0.982]$	$[0.991, 0.983]$
(b)	DWT	$[0.974, 0.989]$	$[0.980, 0.992]$
	DWPT	$[0.985, 0.971]$	$[0.990, 0.976]$
(c)	DWT	$[0.927, 0.977]$	$[0.932, 0.964]$
	DWPT	$[0.944, 0.968]$	$[0.952, 0.952]$

A measure of classification ability was a fraction of correct classifications to the background and to the object, before $[P_1, P_2]$ and after the refinement $[P_{1r}, P_{2r}]$. The results are given in Table 2.

8 Conclusions

Our experiments have shown that the DWPT outperforms the DWT in the ability of generating better features for classification. However, the worst result was achieved for the Fig. 2c, where the object of interest was composed of many sub-objects of not enough size for calculating good texture features.

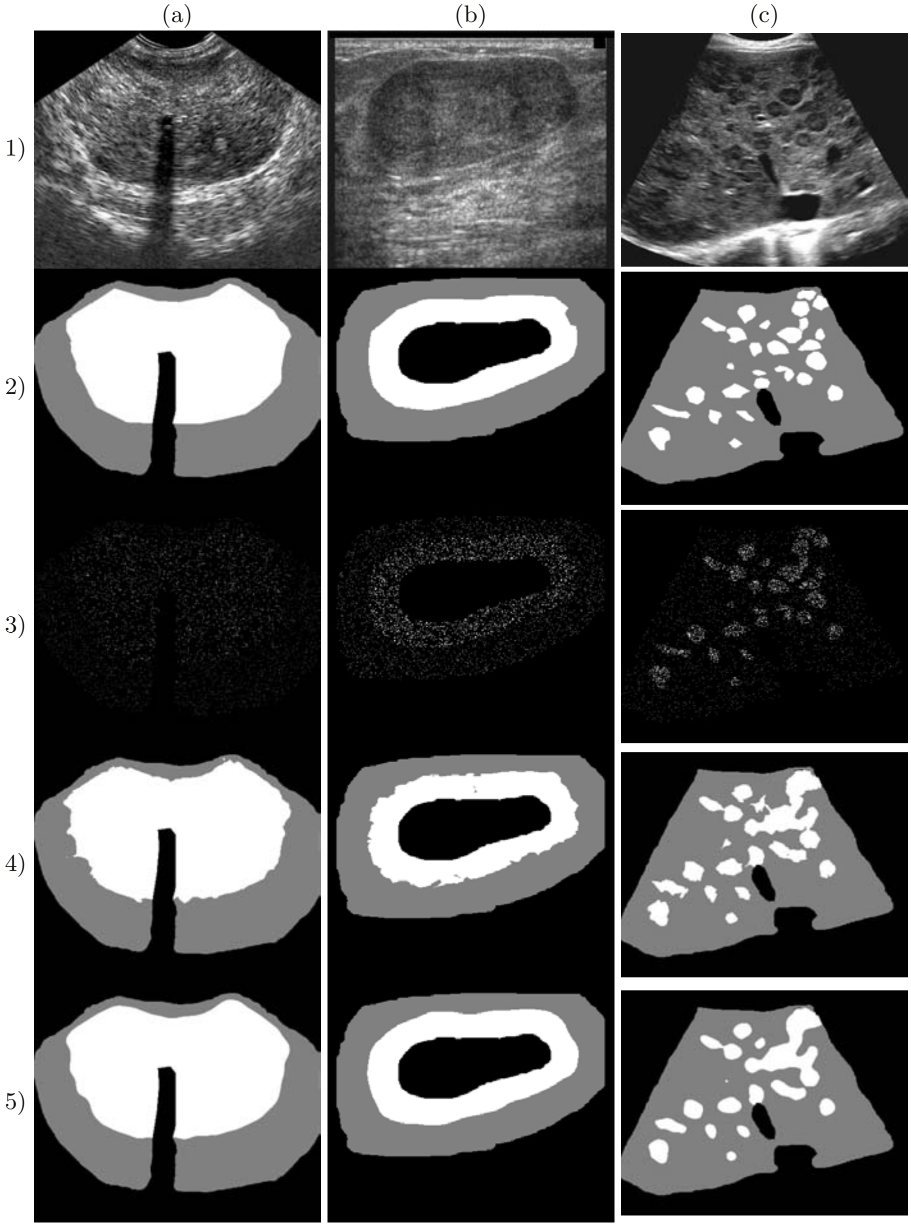


Fig. 2. Ultrasound images used in experiments: (a) a transrectal section of prostate, (b) a solid breast mass, (c) liver metastases. The rows contains: 1) input image, 2) correct pixel classification with ROI, 3) training set, 4) raw results of classification, 5) refined results by using a neighbourhood. Used colours: gray – the background (class #1), white – the object (class #2), black – the omitted area.

The use of neighbourhood information improves classification results only for the images (a) and (b). It decreases classification correctness for object's pixels in the image (c). We suppose that this is caused by an incorrect choice of the Gaussian filter for this image.

Of course, the proposed ROI reduction method can be used in time constrained applications only. In application where an accuracy is the main requirement one should not reduct the training set in a purely probabilistic way.

We suppose that a transformation of the feature space reducing the number of features should facilitate the training and classification process, e.g. the Karhunen-Loeve transform, whose usefulness for various textures was proved [10].

Acknowledgements

This material is based upon work supported by the grant 4 T11E 017 22 from the Polish State Committee for Scientific Research in 2002-05.

References

1. Arambula Cosio F.,Davies B.L., "Automated prostate recognition: a key process for clinically effective robotic prostatectomy", *Med. & Biol. Eng. & Comput.*, Vol. 37, pp. 236-243, 1999.
2. Daubechies I., *Ten Lectures on Wavelets*, Philadelphia: Soc. Ind. Appl. Math., 1992.
3. Duda R.O., Hart P.E., Stork D.G., *Pattern Classification*, John Wiley & Sons, 2001.
4. Haralick R.M., "Statistical and structural approaches to texture", *Proc. IEEE*, Vol. 67, pp. 786-804, 1979.
5. Jain A.K., Dubes R.C., *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, NJ, 1988.
6. Kieś P., "On Application of Wavelet Transforms to Segmentation of Ultrasound Images", *Proc. of Intern. Conf. on Computer Vision and Graphics*, in print, 2004.
7. Laine A., Fan J., "An Adaptive Approach for Texture Segmentation by Multichannel Wavelet Frames", Center for Computer Vision and Visualization, TR-93-025, 1993.
8. Liu J.-F., Lee J. Ch.-M., "An Efficient and Effective Texture Classification Approach Using a New Notion in Wavelet Theory", *Proc. of ICPR'96*, pp. 820-824, IEEE, 1996.
9. Randen T., *Filter and Filter Bank Design for Image Texture Recognition*, Doctoral dissert., Norwegian Univ. of Science and Techn., Stravanger College, 1997.
10. Unser M., Eden M., "Multiresolution Feature Extraction and Selection for Texture Segmentation", *IEEE Trans. on Pattern Anal. & Mach. Intell.*, Vol. 11, No. 7, pp. 717-728, July 1989.
11. Unser M., "Texture Classification and Segmentation Using Wavelet Frames", *IEEE Trans. on Image Processing*, Vol. 4, No. 11, pp. 1549-1560, Nov. 1995.
12. Vapnik V.N., *The Nature of Statistical Learning Theory*, Springer, New York, 1995.

Use of Neural Networks in Automatic Caricature Generation: An Approach Based on Drawing Style Capture

Rupesh N. Shet, Ka H. Lai, Eran A. Edirisinghe, and Paul W.H. Chung

Department of Computer Science, Loughborough University, UK
rupesh_shet@hotmail.com

Abstract. Caricature is emphasizing the distinctive features of a particular face. Exaggerating the Difference from the Mean (EDFM) is widely accepted among caricaturists to be the driving factor behind caricature generation. However the caricatures created by different artists have different drawing style. No attempt has been taken in the past to identify these distinct drawing styles. Yet the proper identification of the drawing style of an artist will allow the accurate modelling of a personalised exaggeration process, leading to fully automatic caricature generation with increased accuracy. In this paper we provide experimental results and detailed analysis to prove that a Cascade Correlation Neural Network (CCNN) can be used for capturing the drawing style of an artist and thereby used in realistic automatic caricature generation. This work is the first attempt to use neural networks in this application area and have the potential to revolutionize existing automatic caricature generation technologies.

1 Introduction

Caricature is an art that conveys humour to people via drawing human faces. The basic concept is capturing the essence of a persons face by graphically exaggerating their distinctive facial features. Many approaches have been proposed in literature to generate facial caricatures automatically [1-4]. The process of creating a caricature even a professional caricaturist would not be able to quantify all the exaggerations he/she is likely to introduce. It is observed that these exaggerations often depend on the individual drawing style adopted by an artist. The fact that we are able to identify caricatures drawn by famous caricaturists, regardless of the original image, supports this observation. Unfortunately none of the existing state-of-the-art automatic caricature generation techniques attempt to capture the drawing style of an individual artist. Yet the accurate capture of this detail would allow more realistic caricatures to be generated. From the artists' point of view, it is difficult for them to explain how they draw caricatures. This is because the drawing rules are embedded in their subconscious mind and often unexplainable. Automatic identification of an artist's drawing style using artificial intelligence techniques could provide a solution for this.

The human brain has an innate ability of remembering and recognising thousands of faces it encounters during a lifetime. Psychologists [3,4] suggested that human beings have a "mean face" recorded in their brain, which is an *average* of faces they encounter in life. A caricaturist compares one's face with this mean face and draws caricatures by exaggerating the distinctive facial features. This caricature drawing

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 343–351, 2005.

approach is widely accepted among psychologists and caricaturists [1,5]. Within the wider aspect of our research we are currently investigating the full automation of the above mentioned drawing style capture and related caricature generation process. The work presented in this paper limits the investigation to capturing the drawing style adopted by a caricaturist in exaggerating a single, selected facial component. Capturing the drawing style of a complete face is a challenging task due to the large number of possible variations and non-linearity of exaggerations that a caricaturist may adopt for different facial components. However non-linearity in exaggerations could be found even in the deformations made to a single facial component. This observation undermines previous research, which assumes semi-linear deformations over a single facial component such as an eye, mouth, nose etc. Fortunately neural networks have the ability to capture the non-linear relationship. Within the research context of this paper we provide experimental results and analysis to prove that a Cascade Correlation Neural Network (CCNN) [8,9] can be trained to accurately capture the drawing style of a caricaturist in relation to an individual facial object. Further we use the results to justify that the trained CCNN could then be used to automatically generate a caricature (drawn by the same artist) of the same facial component belonging to either the same original facial figure or of a different one.

This paper is organised as follows: section-2 introduces the CCNN and discusses its suitability for the application domain. Section-3 presents the proposed methodology of using CCNN in identifying the drawing style of an artist. Section-4 presents experimental results and a detailed analysis proving the validity of the proposed concepts use in capturing the drawing style of an artist. Finally section-5 concludes with an insight into further research that is currently being considered as a result of it.

2 The Cascade Correlation Neural Network

Artificial neural networks are the combination of artificial neurons that are widely used in machine learning. After testing and analysing various neural networks we found that the CCNN is the best for the application domain under consideration.

The CCNN [8,9] is a new architecture and is a generative, feed forward, supervised learning algorithm for artificial neural networks. It is similar to a traditional network in which the neuron is the most basic unit. However an untrained CCNN will remain in a blank state with no hidden units. A hidden neuron is ‘recruited’ when training yields no appreciable reduction of error. Thus a pool of hidden neurons is created with a mixture of non-linear activation functions. The resulting network is trained until the error reduction halts. The hidden neuron with the greatest correspondence to the overall error is then installed in the network and the others are discarded. The new hidden neuron ‘rattles’ the network and significant error reduction is accomplished after each inclusion. The features they identify are permanently cast into the memory of the network, which means that it has the ability to detect the features from training samples. Preserving the orientation of hidden neurons allows cascade correlation to accumulate experience after its initial training session. The above features justify its use within the application domain under consideration. In addition the CCNN has several other advantages [8] namely: 1) It learns very quickly and is at least ten times faster than traditional back-propagation algorithms. 2) The network determines its own size and topology and retains the structure. 3) It is useful for incremental learning in which new information is added to the already trained network.

Once the architecture has been selected and the input signals have been prepared, the next step is to train the neural network. We use the Levenberg-marquardt back-propagation training function [9] due to its significant speed of operation. After training, the accuracy and capabilities of this trained neural network has been validated before putting into practise. In section 4 we validate the use of the above network within the application domain under consideration.

3 Capturing the Drawing Style of a Caricaturist: The Proposed Methodology

Figure 1 illustrates the block diagram of the proposed drawing style capture algorithm. A facial component extractor module subdivides a given original facial image, its corresponding caricature drawn by the artist and the mean face into distinguishable components such as eye, nose etc. Then the geometrical data from a given component of an original image and data from the corresponding component of the mean image are entered as inputs to the neural network module. The relevant data from the caricature component is entered to the module as the output. The above data is used to train the neural network. Once sufficient data points have been used in the above training process, we show that the neural network is able to predict the caricature of a novel image depicting the same facial component that was used in the training process.

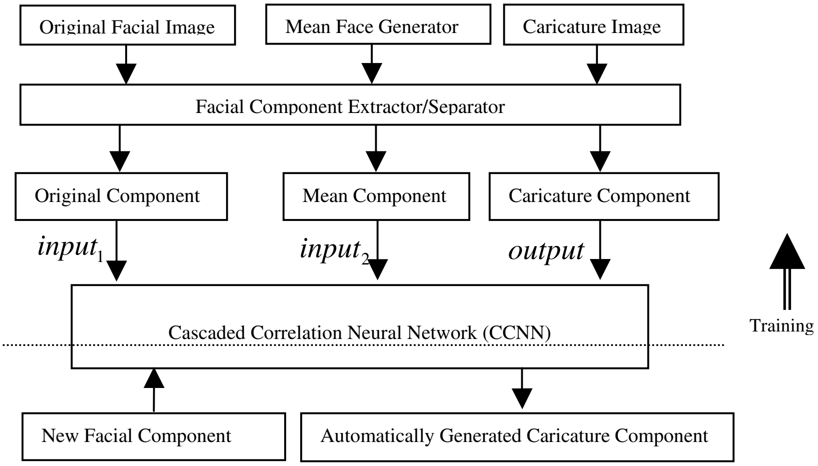


Fig. 1. Proposed Drawing Style Capture Algorithm

Step 1: Generating Mean Face: For the purpose of our present research which is focused only on a proof of concept, the mean face (and thus the facial components) was hand drawn for experimental use and analysis. However, in a real system one could use one of the many excellent mean face generator programs [12].

Step 2: Facial Component Extraction/Separation: A simple image analysis tool based on edge detection, thresholding and thinning was developed to extract/separate various significant facial components such as, ears, eyes, nose and mouth from the original, mean and caricature facial images (see figure 2). Many such algorithms and commercial software packages are available to extract facial components [11].

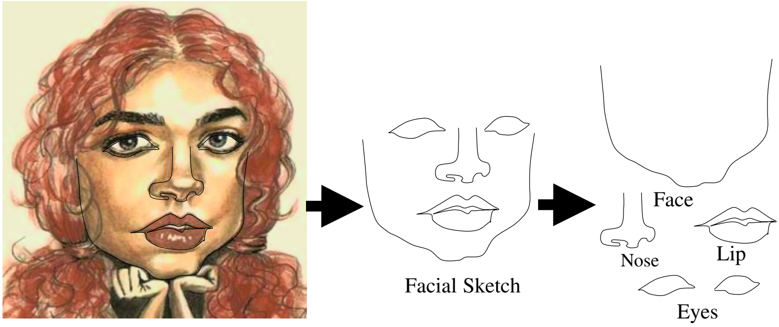


Fig. 2. Facial component extraction from a caricature image

Step 3: Creating Data Sets for Training the Neural network: Once the facial components have been extracted, the original, mean and caricature images of the component under consideration are overlapped, assuming an appropriate common centre point (see figure 3). E.g., for an eye, the centre of the *iris* could be considered the most appropriate centre point. Subsequently using cross sectional lines centred at the above point and placed at equal angular separations, the co-ordinate points at which the lines intersect the components are noted. This is done following a clockwise direction as noted by points 1,2,…8 of the caricature image data set of figure. 3. Note that figure 3 is for illustration purposes only (not to scale) and thus may not represent an accurately scaled/proportioned diagram.

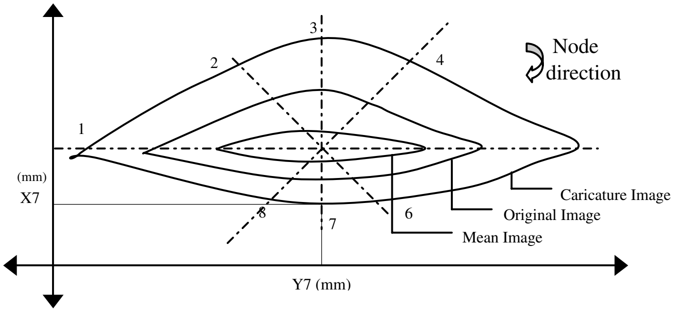


Fig. 3. Creating Data Sets for Training

Step 4: Tabulating Data Sets: After acquiring the X-Y coordinate points as in step-3, they are tabulated as depicted in Table-1.

The higher the number of cross sectional lines that are used, the more accurate the captured shape would be. However for clarity of presentation and ease of experimentation, we have only used four cross sectional lines in figure 3, which results in eight data sets.

Table 1. Training Data Set

	Original	Mean	Caricature		Original	Mean	Caricature	Original
X1	25	37	13		X5	86	75	100
Y1	99	99	99		Y5	99	99	99
X2	47	53	37		X6	62	60	68
Y2	108	102	118		Y6	93	95	87
X3	56.8	56.8	56.8		X7	56.8	56.8	56.8
Y3	109	102	125		Y7	92	95	86
X4	66	59	76		X8	50	52	45
Y4	109	102	119		Y8	93	95	87

Step 5: Data Entry: Considering the fact that the neural network should be trained to automatically produce a caricature of a given facial component drawn by a particular artist, we consider the data points obtained from the caricature image above to be the output training dataset of the neural network. Furthermore the neural network is provided with the data sets obtained from the original and mean images to formulate input data. This follows the widely accepted strategy used by the human brain to analyse a given facial image in comparison to a known mean facial image.

Step 6: Setting up the Neural Network: We propose the use of the following training parameters for a simple, fast and efficient training process.

Table 2. Neural Network Specifications

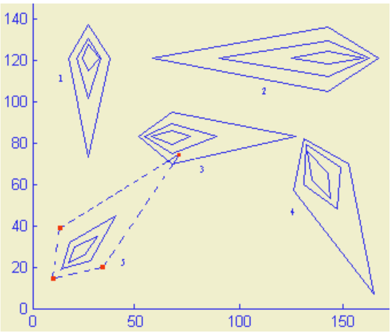
Parameter	Choice
Neural Network Name	Cascade Correlation
Training Function Name	Levenberg-marquardt
Performance Validation Function	Mean squared error
Number of Layers	2
Hidden Layer Transfer Function	Tan-sigmoid with one neuron at the start
Output Layer Transfer Function	Pure-linear with eight neurons

Step 7: Testing: Once training has been successfully concluded as described above, the relevant facial component of a new original image is sampled and fed as input to the trained neural network along with the matching data from the corresponding mean component. In section-4 we provide experimental evidence in support of our proof of concept that a CCNN is able to capture the drawing style of a caricaturist.

4 Experiments and Analysis

Several experiments were designed and carried out to prove the suitability of using a CCNN to capture the drawing style of a caricaturist. The MATLAB neural network toolbox and associated functions [7] were used for the simulations. Two of these core experiments are presented and analysed in detail in this section. Note that experiment 1 use simple geometrical shapes for testing.

Experiment 1: This experiment is designed to prove that CCNN is able to accurately predict orientation, direction and exaggeration. The four training objects denoted by 1-4 in figure 4(a) represent the training cases. In each training object, the innermost shape denotes the mean component, the middle shape denotes the original component and the outermost denotes the caricature component. Note that the exaggeration in one direction is much greater than in the other three directions for all training objects. Object 4 in figure 4(a) denotes the test case. The input shapes (mean and original) are illustrated by continuous lines and the output (i.e. generated caricature) shape is denoted by the dotted shape. Note that the CCNN has been able to accurately predict exaggeration along the proper direction, i.e. along the direction where exaggeration is the most when the original is compared with the mean in the test object.



(a)

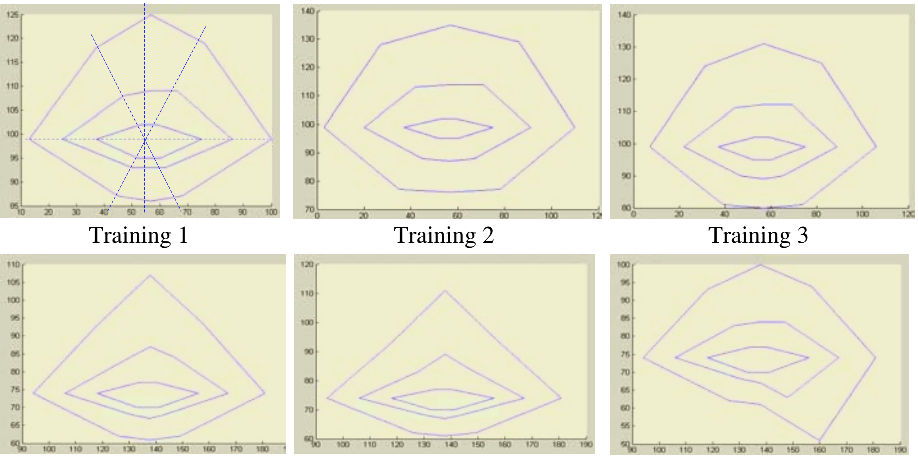
Ob.	X1	Y1	X2	Y2	X3	Y3	X4	Y4
1M	32	121	26.5	127	24	121	26.5	114
1O	21	121	26.5	130	33	121	26.5	101
1C	17	121	26.5	137	37	121	26.5	73
2M	143	124	160	121	143	118	125	121
2O	143	129	163	121	143	112	90	121
2C	143	136	167	121	143	105	58	121
3M	67	79	57	83	67	86	76	83
3O	67	75	54	83	67	89	89	83
3C	67	70	51	83	67	95	127	83
4M	135	62	133	76	143	65.5	144	53
4O	131	60	132	79	149	68	147	48
4C	126	57	131	82	153	70	165	7
5M	25	26	17	22	20	30	31	35
5O	28	23	14	19	18	32	40	45
Test Result								
5C	33	19.2	10.3	15	12.6	38.4	71.27	74.6
(M-mean) (O-Original) (C-Caricature)								

(b)

Fig. 4. Experiment 2 Data (a) graphical (b) tabular

Experiment 2: Experiment 1 was performed on a basic shape and proved that the CCNN is capable of accurately predicting orientation, direction and exaggeration. In this experiment we test CCNN on a more complicated shape depicting a mouth (encloses lower and upper lips). Figure 5 illustrates six training cases out of 20 cases used in the experiment. In each training case, the innermost shape corresponds to a *mean mouth*. For all training and tests cases the shape of the mean mouth has been maintained as constant [Note: To reduce experimental complexity, the sampling points were limited to 8. This does not undermine the experimental accuracy. However more sampling points would have allowed us to train the neural network on a more regular and realistic *mouth* shape.] The middle shape corresponds to the *original mouth* and the outermost shape represents the *caricature mouth*. All these shapes have been sampled at 8 points as illustrated in training case 1 of figure 5. Note the non-linearity in exaggeration that is shown in the training cases across the shape of the mouth. Our set of 20 training cases was carefully selected so as to cover all possible exaggerations in all eight directions. This is a must in order for the CCNN to be able to predict exaggerations accurately in all of the eight directions.

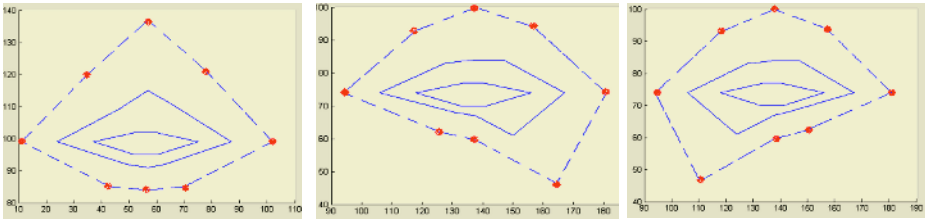
Figures named "result 1-3" in figure 6 below, illustrate the test cases. They demonstrate that the successful training of the CCNN has resulted in its ability to accurately predict exaggeration of non-linear nature in all directions. Note that an increase in the amount of the training data set would result in an increase of the prediction accuracy for a new set of test data.



Training 1 Training 2 Training 3

Training 4 Training 5 Training 6

Fig. 5. Testing CCNN on a real facial object under limited sampling – the training cases



Result1 Result2 Result 3

Fig. 6. Testing CCNN on a real facial object under limited sampling – the test cases

4.1 Analysis: Use of CCNN in Automatic Caricature Generation

Our experiments above were designed to support the proof of concept that the CCNN can be used in capturing the drawing style of an artist and subsequent automatic caricature generation. Here we provide justifications as to why the experiments performed on limited shapes, with limited sampling would still prove enough evidence in support of the proposed idea.

Figure-7 illustrates the mean, original and caricature (drawn by two artists) images of a human eye. The original eye shows a noticeable difference in shape from the mean eye at the two ends. In the left end, the eye is curved up whereas at the right end it is curved down.

The drawing style of artist-1 shows no difference being made to the left side but a noticeable exaggeration to the difference (curved nature) in the right side. This could be a trait of this artist’s drawing style. I.e. the artist makes no exaggerations in any cartoon he draws, in the left corner of the eye, but exaggerates considerably in the right corner. The proposed CCNN based approach is able to learn this rule as proved by the results of experiments 1. Performing experiments on a larger set of original eyes (belonging to different people but caricatured by the same artist-1) will help improve prediction further. Using more sampling points around the surface of the eye (rather than 8 in our experiments) will increase the accuracy of approximating the actual shape of the eye.

In figure 7, the drawing style of artist-2 shows exaggerations being done at both ends of the eye. As justified above and supported by evidence from experiments 2 and 3, CCNN would be capable of accurately capture the drawing style of artist-2 as well. Given a new original eye, it would then be able to automatically generate the caricature, incorporating the artist's style.

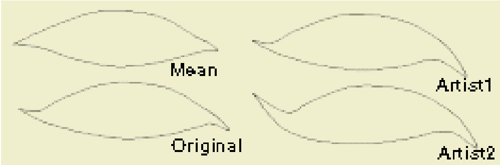


Fig. 7. Comparison of the mean and an original human eye with a caricature eye drawn by two different artists

5 Conclusion

In this paper we have identified an important shortcoming of existing automatic caricature generation systems in that their inability to identify and act upon the unique drawing style of a given artist. We have proposed a Cascade-Correlation Neural Network based approach to identify the said drawing style of an artist by training the neural network on unique non-linear deformations made by an artist when producing caricature of individual facial objects. The trained neural network has been subsequently used successfully to generate the caricature of the facial component automatically. We have shown that the automatically generated caricature consists of various unique straits adopted by the artist in drawing free-hand caricatures.

The above research is a part of a more advanced research project that is looking at fully automatic, realistic, caricature generation of complete facial figures. One major challenge faced by this project includes, non-linearities and unpredictabilities of deformations introduced in exaggerations done between different objects within the same facial figure, by the same artist. We are currently extending the work of this paper in combination with artificial intelligence technology to find an effective solution to the above problem.

References

1. S. E. Brennan.: Caricature Generator: The Dynamic Exaggeration of Faces by Computer, Leonardo, Vol. 18, No. 3. (1985) 70-178
2. P.J. Benson, D.I. Perrett.: Synthesising Continuous-tone Caricatures, Image & Vision Computing, Vol. 9. (1991) 123-129
3. G. Rhodes, T. Tremewan.: Averageness, Exaggeration and Facial Attractiveness, Psychological Science, Vol. 7. (1996) 105-110
4. J.H. Langlois, L.A. Roggman, L. Mussleman.: What Is Average and What Is Not Average About Attractive Faces, Psychological Science, Vol. 5. (1994) 214-220

5. L. Redman.: How to Draw Caricatures, McGraw-Hill Publishers (1984)
6. Neural Network.: http://library.thinkquest.org/C007395/tqweb/index.html (Access date 13th Oct. 2004)
7. The Maths Works Inc, User's Guide version 4, Neural Network Toolbox, MATLAB
8. S. E. Fahlman.: The Cascade-Correlation Learning Architecture, Technical Report CMU-CS-90-100, School of Computer Science, Carnegie Mellon University (1990)
9. Carling, A.: Introducing Neural Networks, Wilmslow, UK, Sigma Press (1992)
10. Fausett, L.: Fundamentals of Neural Networks, New York, Prentice Hall (1994)
11. http://www.jasc.com/products (Access date 11th Nov. 2004)
12. http://www.asahi-net.or.jp/~FX6M-FJMY/mop00e.html M.Fujimiya. "Morpher" (Access date 11th Nov. 2004)

Part IV

Document Analysis

Information Theoretic Text Classification Using the Ziv-Merhav Method

David Pereira Coutinho[1] and Mário A.T. Figueiredo[2]

[1] Depart. de Engenharia de Electrónica e Telecomunicações e de Computadores
Instituto Superior de Engenharia de Lisboa
1959-007 Lisboa, Portugal
davidpc@isel.pt

[2] Instituto de Telecomunicações
Instituto Superior Técnico
1049-001 Lisboa, Portugal
mtf@lx.it.pt

Abstract. Most approaches to text classification rely on some measure of (dis)similarity between sequences of symbols. Information theoretic measures have the advantage of making very few assumptions on the models which are considered to have generated the sequences, and have been the focus of recent interest. This paper addresses the use of the *Ziv-Merhav method* (ZMM) for the estimation of relative entropy (or Kullback-Leibler divergence) from sequences of symbols as a tool for text classification. We describe an implementation of the ZMM based on a modified version of the Lempel-Ziv algorithm (LZ77). Assessing the accuracy of the ZMM on synthetic Markov sequences shows that it yields good estimates of the Kullback-Leibler divergence. Finally, we apply the method in a text classification problem (more specifically, authorship attribution) outperforming a previously proposed (also information theoretic) method.

1 Introduction

Defining a similarity measure between two finite sequences, without explicitly modelling their statistical behavior, is a fundamental problem with many important applications in areas such as information retrieval or text classification. Approaches to this problem include: various types of edit (or Levenshtein) distances between pairs of sequences (*i.e.*, the minimal number of edit operations, chosen from a fixed set, required to transform one sequence into the other; see, *e.g.*, [1], for a review); "universal" distances (*i.e.* independent of a hypothetical source model) such as the *information distance* [2]; methods based on universal (in the Lempel-Ziv sense) compression algorithms [3].

In this paper, we consider using the method proposed by Ziv and Merhav (ZM) for the estimation of relative entropy, or Kullback-Leibler (KL) divergence, from pairs of sequences of symbols, as a tool for text classification. In particular, to handle the text authorship attribution problem, Benedetto, Caglioti and

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 355–362, 2005.

Loreto [3] introduced a "distance" function based on an estimator of the relative entropy obtained by using the *gzip* compressor [4] and file concatenation. This work follows the same idea of estimating a dissimilarity using data compression, but using the ZM method [5]. The ZM approach avoids the drawbacks of the method of Benedetto *et al* [3] which have been pointed out by Puglisi *et al* [6], and has desirable theoretical properties of fast convergence.

We describe an implementation of the ZM method based on a modified version of the Lempel-Ziv algorithm. We assess the accuracy of the ZM estimator on synthetic Markov sequences, showing that it yields good estimates of the KL divergence. Finally, we apply the method to an authorship attribution problem using a text corpus similar to the one used in [3]. Our results show that ZM method outperforms the technique introduced in [3].

The outline of the paper is has follows. In Section 2 we recall the fundamental tools used in this approach: the concept of relative entropy, the method proposed by Bennedeto *et al*, and the ZM method. In Section 3 we describe our implementation of the ZM technique based on the LZ77 algorithm. Section 4 presents the experimental results, while Section 5 concludes the paper.

2 Data Compression and Similarity Measures

2.1 Kullback-Leibler Divergence and Optimal Coding

Consider two memoryless sources $\mathcal{A}$ and $\mathcal{B}$ producing sequences of binary symbols. Source $\mathcal{A}$ emits a 0 with probability p (thus a 1 with probability $1-p$) while $\mathcal{B}$ emits a 0 with probability q. According to Shannon [7, 8], there are compression algorithms that applied to a sequence emitted by $\mathcal{A}$ will be asymptotically able to encode the sequence with an average number bits per character equal to the source entropy $H(\mathcal{A})$, *i.e.*, coding, on average, every character with

$$H(\mathcal{A}) = -p\,\log_2 p - (1-p)\,\log_2 (1-p) \quad \text{bits.} \tag{1}$$

An optimal code for $\mathcal{B}$ will not be optimal for $\mathcal{A}$ (unless, of course, $p = q$). The average number of extra bits per character which are wasted when we encode sequences emitted by $\mathcal{A}$ using an optimal code for $\mathcal{B}$ is given by the relative entropy (KL divergence) between $\mathcal{A}$ and $\mathcal{B}$ (see, *e.g.*, [8]), that is

$$D(\mathcal{A}||\mathcal{B}) = p \log_2 \frac{p}{q} + (1-p) \log_2 \frac{1-p}{1-q}. \tag{2}$$

This fact suggests the following possible way to estimate the KL divergence between two sources: design an optimal code for source $\mathcal{B}$ and then measure the average number of bits obtained when this code is used to encode sequences from source $\mathcal{A}$. The difference between this average code length and the entropy of $\mathcal{A}$ is an estimate of the KL divergence $D(\mathcal{A}||\mathcal{B})$. The entropy of $\mathcal{A}$ itself can be estimated by measuring the average code length of an adapted optimal code. This is the basic idea that underlies the methods proposed in [3] and [5]. However, to use this idea for general sources (not simply for the memoryless ones that we have considered up to now for simplicity), without having to explicitly estimate models for each of them, we need to use some form of universal coding. A universal coding technique (such as the Lempel-Ziv algorithm) is one that is asymptotically able to achieve the entropy lower bound without prior knowledge of the source distribution (which, of course, does not have to be memoryless) [8].

2.2 Relationship Between Entropy and Lempel-Ziv Coding

Consider a sequence $\mathbf{x} = (x_1, x_2, ..., x_n)$ emitted by an unknown lth-order stationary Markovian source, defined over a finite alphabet. Suppose that one wishes to estimate the nth-order entropy, or equivalently $-(1/n)\log_2 p(x_1, x_2, ..., x_n)$. A direct approach to this goal is computationally prohibitive for large l, or even impossible if l is unknown. However, an alternative route can be taken using the following fact (see [8], [9]): the Lempel-Ziv (LZ) code length for $\mathbf{x}$, divided by n, is a computationally efficient and reliable estimate of the entropy, and hence also of $-(1/n)\log_2 p(x_1, x_2, ..., x_n)$. More formally, let $c(\mathbf{x})$ denote the number of phrases in $\mathbf{x}$ resulting from the LZ sequential parsing of $\mathbf{x}$ into distinct phrases, such that each phrase is the shortest sequence which is not a previously parsed phrase. Then, the LZ code length for $\mathbf{x}$ can be approximated by

$$c(\mathbf{x})\log_2 c(\mathbf{x}) \tag{3}$$

and it can be shown that it converges almost surely to $-(1/n)\log_2 p(x_1, x_2, ..., x_n)$, as $n \to \infty$ [5]. This shows that we can use the output of an LZ encoder to estimate the entropy of an unknown source without explicitly estimating its model parameters.

2.3 The Method of Benedetto, Caglioti and Loreto

Recently, Benedetto *et al* [3] have proposed a particular way of using LZ coding to estimate KL divergence between two sources $\mathcal{A}$ and $\mathcal{B}$. They have used the proposed method for context recognition and classification of sequences.

Let $|X|$ denote the length in bits of the uncompressed sequence X, let L_X denote the length in bits obtained after compressing sequence X (in particular, [3] uses *gzip*, which is an LZ-based compression algorithm [4]), and let $X + Y$ stand for the concatenation of sequences X and Y (with Y after X). Let A and B be "long" sequences from sources $\mathcal{A}$ and $\mathcal{B}$, respectively, and b a "small" sequence from source $\mathcal{B}$. As proposed by Benedetto *et al*, the relative entropy $D(\mathcal{A}||\mathcal{B})$ (per character) can be estimated by

$$\widehat{D}(\mathcal{A}||\mathcal{B}) = (\Delta_{Ab} - \Delta_{Bb})/|b|, \tag{4}$$

where $\Delta_{Ab} = L_{A+b} - L_A$ and $\Delta_{Bb} = L_{B+b} - L_B$. Notice that $\Delta_{Ab}/|b|$ can be seen as the code length (per character) obtained when coding a sequence from $\mathcal{B}$ (sequence b) using a code optimized for $\mathcal{A}$, while $\Delta_{Bb}/|b|$ can be interpreted as an estimate of the entropy of the source $\mathcal{B}$.

To handle the text authorship attribution problem, Benedetto, Caglioti and Loreto (BCL) [3] defined a simplified "distance" function $d(A, B)$ between sequences,

$$d(A, B) = \Delta_{AB} = L_{A+B} - L_A, \tag{5}$$

which we will refer to as the BCL divergence. As mention before, Δ_{AB} is a measure of the description length of B when the coding is optimized to A, obtained by subtracting the description length of A from the description length of $A + B$. Hence, it can be stated that $d(A, B'') < d(A, B')$ means that B'' is more similar to A than B'. Notice that the BCL divergence is not symmetric.

More recently, Puglisi *et al* [6] studied in detail what happens when a compression algorithm, such as LZ77 [10], tries to optimize its features at the interface between two different sequences A and B, while compressing the sequence $A + B$. After having compressed sequence A, the algorithm starts compressing sequence B using the dictionary that it has learned from A. After a while, however, the dictionary starts to become adapted to sequence B, and when we are well into sequence B the dictionary will tend to depend only on the specific features of B. That is, if B is long enough, the algorithm learns to optimally compress sequence B. This is not a problem when the sequence B is so short that the dictionary does not become completely adapted to B. In this case, one can measure the relative entropy by compressing the sequence $A+B$. The problem arises for long sequences B. The Ziv-Merhav method, described next, does not suffer from this problem, this being what motivated us to consider it for sequence classification problems.

2.4 Ziv-Merhav Empirical Divergence

The method proposed by Ziv and Merhav [5] for measuring relative entropy is also based on two Lempel-Ziv-type parsing algorithms:

- The incremental LZ parsing algorithm [9], which is a self parsing procedure of a sequence into $c(\mathbf{z})$ distinct phrases such that each phrase is the shortest sequence that is not a previously parsed phrase. For example, let $n = 11$ and $\mathbf{z} = (01111000110)$, then the self incremental parsing yields $(0, 1, 11, 10, 00, 110)$, namely, $c(\mathbf{z}) = 6$.
- A variation of the LZ parsing algorithm described in [5], which is a sequential parsing of a sequence $\mathbf{z}$ with respect to another sequence $\mathbf{x}$ (cross parsing). Let $c(\mathbf{z}|\mathbf{x})$ denote the number of phrases in $\mathbf{z}$ with respect to $\mathbf{x}$. For example, let $\mathbf{z}$ as before and $\mathbf{x} = (10010100110)$; then, parsing $\mathbf{z}$ with respect to $\mathbf{x}$ yields $(011, 110, 00110)$, that is $c(\mathbf{z}|\mathbf{x}) = 3$.

Ziv and Merhav have proved that for two finite order (of any order) Markovian sequences of length n the quantity

$$\Delta(\mathbf{z}||\mathbf{x}) = \frac{1}{n}\left[\, c(\mathbf{z}|\mathbf{x}) \, \log_2 n - c(\mathbf{z}) \, \log_2 c(\mathbf{z}) \,\right] \tag{6}$$

converges, as $n \to \infty$, to the relative entropy between the two sources that emitted the two sequences $\mathbf{z}$ and $\mathbf{x}$. Roughly speaking, we can observe (see (3)) that $c(\mathbf{z}) \log_2 c(\mathbf{z})$ is the measure of the complexity of the sequence $\mathbf{z}$ obtained by self-parsing, thus providing an estimate of its entropy, while $(1/n)\, c(\mathbf{z}|\mathbf{x}) \log_2 n$ can be seen as an estimate of the code-length obtained when coding $\mathbf{z}$ using a model for $\mathbf{x}$. From now on we will refer to $\Delta(\mathbf{z}||\mathbf{x})$ as the ZM divergence.

3 Modified LZ77 Algorithm

We have implemented the ZM divergence using the LZ78 algorithm to make the self parsing procedure. To perform the cross parsing, we designed a modified LZ77-based algorithm where the dictionary is static and only the lookahead buffer slides over the input sequence. For better understanding, let us briefly recall the LZ77 algorithm and its implementation model.

The LZ77 compression algorithm observes the input sequence through a sliding window buffer as shown in Figure 1. The sliding window buffer consists of a dictionary and a *lookahead buffer* (LAB). The dictionary holds the symbols already analyzed and the LAB the symbols to be analyzed. At each step, the algorithm tries to express the sequence in the LAB as a subsequence in the dictionary using a reference to it and then coding that match. Otherwise, the leftmost symbol in the LAB is coded as a literal. In both situations, the dictionary is updated after each step.

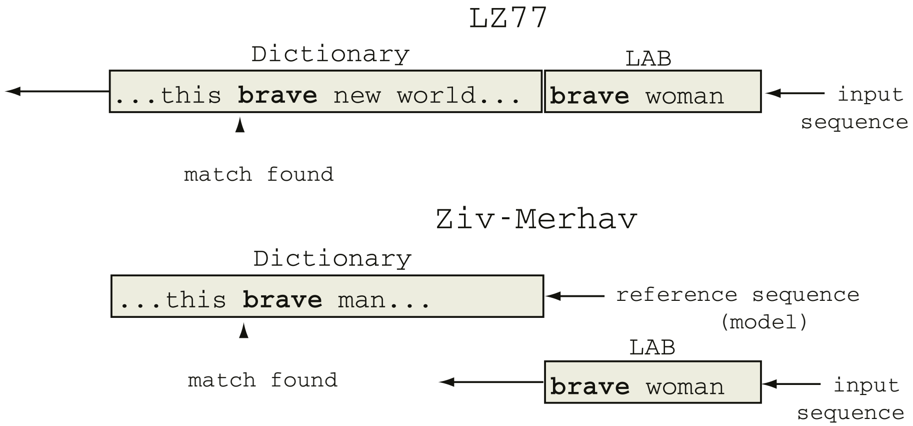


Fig. 1. The original LZ77 algorithm uses a sliding window over the input sequence to get the dictionary updated, whereas in the Ziv-Merhav cross parsing procedure the dictionary is static and only the *lookahead buffer* (LAB) slides over the input sequence.

To implement the cross parsing procedure, we first use the reference sequence (model) to build an LZ77-like dictionary, which will remain static. After that, the input sequence (to be compared) slides through the LAB from right to left as shown in Figure 1. At each step, the procedure is the same as with LZ77, except that the dictionary is not updated.

Two important parameters of the algorithm are the dictionary size and the maximum length of a matching sequence found in the LAB; both influence the parsing results and determine the compressor efficiency [4]. The experiments reported in the next section were performed using a 65536 byte dictionary and a 256 byte long LAB.

4 Experiments

4.1 Synthetic Data

The purpose of our first experiments was to compare the theoretical values of the KL divergence with the estimates produced by the ZM method, on pairs of binary sequences with 100, 1000 and 10000 symbols. The sequences were randomly generated from simulated sources using memoryless and order-1 Markov models. For the memoryless sources, the KL divergence is given by expression (2), while for the order-1 sources it is given by

$$D(p||q) = \sum_{x_1, x_2} p(x_1, x_2) \, \log_2 \frac{p(x_2|x_1)}{q(x_2|x_1)}. \tag{7}$$

Results for these experiments are shown in Figure 2. Each experiment compares KL divergence against ZM divergence, over a varying range of source symbol probabilities. The results show that the ZM divergence provides a good KL divergence estimate, regardless its negative values when the sequences are very similar or "close".

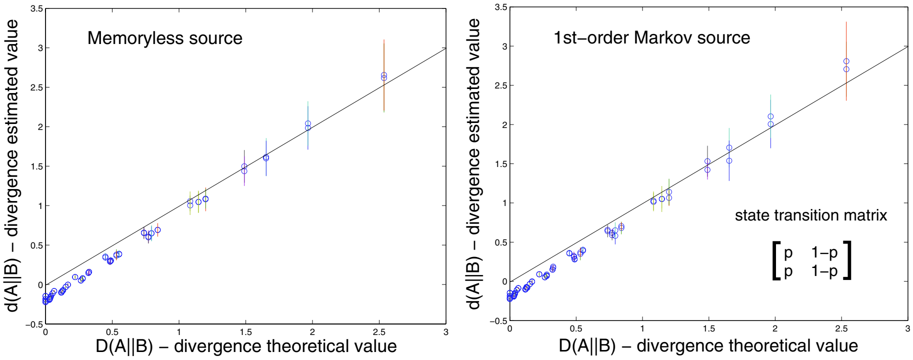


Fig. 2. Theoretical values versus Ziv-Merhav empirical divergence values, between two synthetic binary sequences of 10000 symbols length. Each circle is the sample mean value and the vertical segments are the sample standard deviation values, evaluated over 100 sequence pairs. For the 1st-order Markov source we use the state transition matrix shown and test for all probabilities $p \in [0, 1]$. Results are near to the identity line of no estimation error.

4.2 Text Classification

Our next step was to compare the performance of ZM divergence with the BCL divergence on the authorship attribution problem using a text corpus similar to the one used by Benedetto *et al* [3]. For this purpose, we have used a set of 86 files of the same authors, downloaded from the same site: `www.liberliber.it`. Since we don't know exactly which files were used in [3], we apply both measures to this new corpus of Italian authors. In this experiment, each text is classified as belonging to the author of the closest text in the remaining set. In other words, the results reported can be seen as a full *leave-one-out cross-validation* (LOO-CV) performance measure of a nearest-neighbor classifier built using the considered divergence functions.

Table 1. Italian Authors Classification - For each author we report the number of texts considered and two measures of classification success, one obtained using the original method proposed by Benedetto, Caglioti and Loreto (BCL) and the other with the Ziv-Merhav method (ZM).

author	**No. of texts**	**BCL**	**ZM**
Alighieri	8	7	7
Deledda	15	15	15
Fogazzaro	5	3	5
Guicciardini	6	6	5
Macchiavelli	12	11	11
Manzoni	4	4	3
Pirandello	11	9	11
Salgari	11	11	11
Svevo	5	5	5
Verga	9	7	9
Total	**86**	**78**	**82**

The results of this experiment, which are presented in Table I, show that the ZM divergence outperforms the BCL divergence over the very same corpus. Our rate of success using the ZM divergence is 95.4%, while the BCL divergence achieves rate of success of 90.7%.

5 Conclusion

We have presented an implementation of the Ziv-Merhav method for the estimation of relative entropy or Kullback-Leibler divergence from sequences of symbols, which can be used as a tool for text classification. Computational experiments showed that this method yields good estimates of the relative entropy on synthetic Markov sequences. Moreover, this method was applied to a text classification problem (authorship attribution), outperforming a previously proposed approach. Future work will include further experimental evaluation of the Ziv-Merhav method, as well as its use in more sophisticated text classification algorithms such as a kernel-based methods [11].

References

1. D. Sankoff and J. Kruskal, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison.* Reading, MA: Addison-Wesley, 1983.
2. C. Bennett, P. Gacs, M. Li, P. Vitanyi, and W. Zurek, "Information distance," *IEEE Transactions on Information Theory*, vol. 44, pp. 1407–1423, 1998.
3. D. Benedetto, E. Caglioti, and V. Loreto, "Language trees and zipping," *Physical Review Letters, 88:4*, 2002.
4. M. Nelson and J. Gailly, *The Data Compression Book 2nd edition.* M&T Books, New York, 1995.
5. J. Ziv and N. Merhav, "A measure of relative entropy between individual sequences with application to universal classification," *IEEE Trans. on Information Theory, pp. 1270–1279*, 1993.
6. A. Puglisi, D. Benedetto, E. Caglioti, V. Loreto, and A. Vulpiani, "Data compression and learning in time sequences analysis," *Physica D 180, 92*, 2003.
7. C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal,27: pp. 379-423, pp. 623-656*, 1948.
8. T. Cover and J. Thomas, *Elements of Information Theory.* John Wiley & Sons, Inc, 1991.
9. J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Transactions on Information Theory*, vol. 24, no. 5, pp. 530–536, 1978.
10. J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Transactions on Information Theory*, vol. 23, no. 3, pp. 337–343, 1977.
11. J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Recognition.* Cambridge University Press, 2004.

Spontaneous Handwriting Text Recognition and Classification Using Finite-State Models*

Alejandro Héctor Toselli, Moisés Pastor, Alfons Juan, and Enrique Vidal

Instituto Tecnológico de Informática and
Departamento de Sistemas Informáticos y Computación,
ITI/DSIC, Universidad Politécnica de Valencia, 46071 Valencia, Spain
{ahector,moises,ajuan,evidal}@iti.upv.es

Abstract. Finite-state models are used to implement a handwritten text recognition and classification system for a real application entailing casual, spontaneous writing with large vocabulary. Handwritten short phrases which involve a wide variety of writing styles and contain many non-textual artifacts, are to be classified into a small number of predefined classes. To this end, two different types of statistical framework for phrase recognition-classification are considered,based on finite-state models. HMMs are used for text recognition process. Depending to the considered architecture, N-grams are used for performing text recognition and then text classification (serial approach) or for performing both simultaneously (integrated approach). The multinomial text classifier is also employed in the classification phase of the serial approach. Experimental results are reported which, given the extreme difficulty of the task, are encouraging.

1 Introduction

Cursive handwritten text recognition is currently becoming increasingly mature, thanks to the introduction of holistic approaches based on segmentation-free image recognition technologies. Using these techniques, very good results have been reported for applications entailing relatively clean and homogeneous handwriting and small vocabularies [1–5]. However, as the writing style becomes increasingly variable and spontaneous and/or the number of words to be recognized grows, performance of these systems tends to degrade dramatically.

Here we consider a handwritten text recognition and classification application entailing casual, spontaneous writing and a relatively large vocabulary. In this application, however, the extreme difficulty of text recognition is somehow compensated by the simplicity of the target result. The application consists of classifying (into a small number of predefined classes) casual handwritten answers extracted from survey forms made for a telecommunication company.[1]

* This work has been supported by *Agencia Valenciana de Ciencia y Tecnología* under contract GRUPOS03/031 and Spanish *Ministerio de Ciencia y Tecnología* under grant TIC2003-08496-C04-02.

[1] Data kindly provided by ODEC, S.A. (www.odec.es)

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 363–370, 2005.

In [6], we proposed to tackle this difficult classification task using a *two-step* or *serial approach*. Using character HMMs integrated with an n-gram language model, recognition is first performed on each handwritten sample; then, the recognized word sequence is classified into one of the given eight classes using a text classifier based also on n-grams. In this work, results under this serial scheme are improved using both n-grams and a multinomial text classifier [7] in the second step. Finally, additional results are reported using a newly proposed holistic recognition-classification scheme.

In the next section, characteristics and difficulties of the handwritten text corpora (from the considered application) are explained. Preprocessing and feature extraction are described in section 3. The adopted probabilistic framework is the topic of section 4. Section 5 considers the models which the system is based on. Experimental results are presented in section 6 and conclusions are drawn in the final section.

2 The Handwritten Phrase Application

The considered application phrases were handwritten by a heterogeneous group of people, without any explicit or formal restriction relative to vocabulary, the resulting application lexicon becomes quite large. On the other hand, since no guidelines are given as to the kind of pen or the writing style to be used, phrase become very variable and noisy. For example, in some samples the stroke thickness is non-uniform and the vertical slant also varies within a sample. Other samples present irregular and non-consistent spacing between words and characters. Also, there are samples written using different case and font types, variable sizes and even including foreign language phrases. On the other hand, noise and non-textual artifacts often appear in the phrases. Among these noisy elements we can find unknown words or words containing orthographic mistakes, as well as underlined and crossed-out words. Unusual abbreviations and symbols, arrows, etc. are also within this category. The combination of these writing-styles and noise may result in partly or entirely illegible samples. Examples of these difficulties are shown in Figure 1.

So far, human operators have been in charge of classifying these phrases. They do it through a fast reading which just aims to grasp the essential meaning of the answers. This implies that not all the words can or need to be perfectly recognized; they just retrieve enough information to get an adequate classification. In particular, the eight classes defined in the application are: telephone rates, coverage problems, mobile telephone problems, customer assistance, customers expressing satisfaction, service complains and generic queries for information. The aim of our system is to help performing this classification as fast and accurately as possible, with a minimal human intervention.

3 Preprocessing and Feature Extraction

The following steps take place in the preprocessing of each text image: noise reduction, line extraction, skew and slant corrections and size normalization.

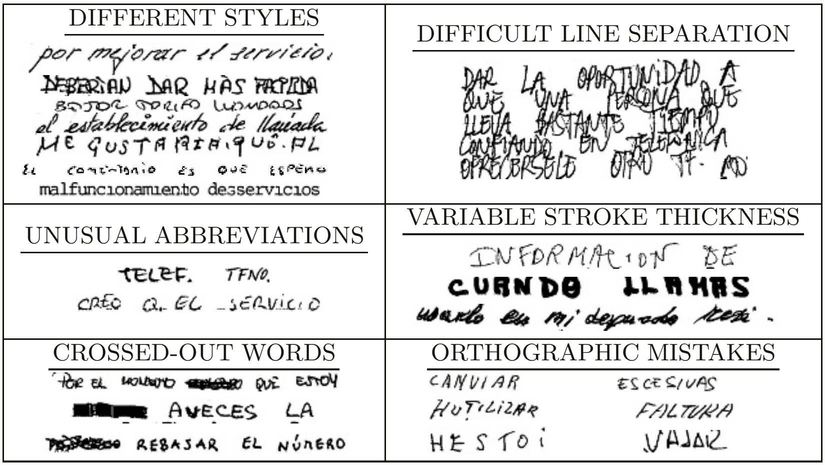


Fig. 1. Some of the difficulties involved in the application.

Because of the inherent difficulty of the task, line extraction is carried out so far in a semi-automatic way, based on a conventional line-extraction method [4]. Most of the phrases are processed automatically, but manual supervision is applied to difficult line-overlapping cases such as that shown in figure 1 (top-right panel). By adequately pasting the extracted lines, a single-line (long) image is obtained.

The skew correction process aims at putting the text line into horizontal position. Detailed information about the process used in this work is described in [5]. On the other hand, the slant correction process applies a horizontal shearing transform to the already deskewed image to bring the writing in an upright position. The method used here, based on projection profile, can be found in [8]. Finally, the size normalization process which tries to make the system invariant to the text height, is detailed in [5].

As with any approach based on (one-dimensional) HMMs, feature extraction must transform the preprocessed image into a *sequence of (fixed-dimension) feature vectors*. To do this, the image is first divided into a grid of small square cells, sized a small fraction of the image height (such as 1/16, 1/20, 1/24 or 1/28). We call this fraction *vertical resolution*. Then each cell is characterized by the following features: *normalized grey level*, *horizontal grey-level derivative* and *vertical grey-level derivative*. To obtain smoothed values of these features, feature extraction is extended to a 5 × 5-cell window centered at the current cell and weighted by a Gaussian function. The derivatives are computed by least squares fitting a linear function. Columns of cells are processed from left to right and a feature vector is built for each column by stacking the features computed in its constituent cells. This process is similar to that followed in [2].

4 Probabilistic Framework

Let $\boldsymbol{x}$ be a sequence of feature vectors extracted from a handwritten line-image and let c identify the meaning of some text (just a classification label, in our case). The ultimately goal of our system is to find an optimal classification for $\boldsymbol{x}$; that is to search for an $\hat{c}$:

$$\hat{c} = \arg\max_{c} P(c \,|\, \boldsymbol{x}) \tag{1}$$

where $P(c \,|\, \boldsymbol{x})$ is the posterior probability that c is the true meaning (class) of $\boldsymbol{x}$. Word recognition is not explicit in this formula, were recognition is seen as a hidden process. Nevertheless, classification can be viewed as a two-step process: $\mathbf{x} \rightarrow \mathbf{s} \rightarrow c$, where s is a sequence of words. To uncover the underlying recognition process, $P(c \,|\, \boldsymbol{x})$ can be seen as a marginal of the joint probability function $P(s, c \,|\, \boldsymbol{x})$. Using the Bayes rule, we can write

$$\hat{c} = \arg\max_{c} \sum_{s} P(s, c \,|\, \boldsymbol{x}) = \arg\max_{c} \sum_{s} p(\boldsymbol{x} \,|\, s, c) P(s, c) \tag{2}$$

$$\approx \arg\max_{c} \sum_{s} p(\boldsymbol{x} \,|\, s) P(s, c) \tag{3}$$

by assuming that, in practice, $p(\boldsymbol{x} \,|\, s, c)$ is independent of c. Approximating the sum by the maximum in eq. (3), we have

$$(\hat{c}, \hat{s}) \approx \arg\max_{c,s} p(\boldsymbol{x} \,|\, s) P(s, c) \tag{4}$$

$$(\hat{c}, \hat{s}) \approx \arg\max_{c,s} p(\boldsymbol{x} \,|\, s) P(s \,|\, c) P(c) \tag{5}$$

were $P(c)$ is the a priori probability of c. Eq. (5) is the basis of our integrated approach to handwriting recognition and classification via finite-state models. This equation permits to simultaneously search for both $\hat{c}$ and its associated most probable decoding, $\hat{s}$. As shown in [5], if $p(\boldsymbol{x} \,|\, s)$, $P(s \,|\, c)$ and $P(c)$ are modeled by finite state models, this problem can be solved efficiently. We adopt conventional HMMs to estimate $p(\boldsymbol{x} \,|\, s)$ (as a sequence of character HMMs) and n-grams to estimate the $P(s \,|\, c)$ of each class c. Thanks to their *homogeneous* finite-state nature, both the HMMs and n-gram classifiers (one for each class c) can be easily integrated into a single *global* finite-state network [4] on which recognition-classification can be efficiently performed.

For the two-step (serial) approach, after replacing $P(s, c)$ by $P(c \,|\, s)P(s)$, eq. (4) is broken down into the following two approximations:

$$\hat{s} \approx \arg\max_{s} p(\boldsymbol{x} \,|\, s) P(s) \tag{6}$$

$$\hat{c} \approx \arg\max_{c} P(c \,|\, \hat{s}) = \arg\max_{c} P(\hat{s} \,|\, c) P(c) \tag{7}$$

For the first (recognition) step (eq. (6)), conventional HMMs which estimate $p(\boldsymbol{x} \,|\, s)$ and n-grams which estimate $P(s)$, are integrated into a single finite-state network [4] on which sentence recognition is efficiently performed. For the second step (eq. (7)), a classifier is built by estimating $P(\hat{s} \,|\, c)$ as an n-gram of words used in the class c. The classification of a recognized word sequence is carried out by directly computing eq. (7) for each class c on the recognized text $\hat{s}$ [9].

Both the integrated approach and the recognition-phase in the two-step approach are done by solving eq. (5) and (6), respectively, using the well known Viterbi algorithm [10]. It is worth nothing that the two-step approximation involves a reduction of the computation demands with respect to integrated approximation. Also, it allows us to use any convenient classification technique for the second step, such as a multinomial Naive Bayes text classifier [7].

5 Character, Word and Sentence Modelling

Sentence models are built by concatenation of *word* models which, in turn, are often obtained by concatenation of continuous left-to-right HMMs for individual *characters*.

Basically, each character HMM is a stochastic finite-state device that models the succession, along the horizontal axis, of (vertical) feature vectors which are extracted from instances of this character. Each HMM state generates feature vectors following an adequate parametric probabilistic law; typically, a *mixture of Gaussian densities*. The required number of densities in the mixture depends, along with many other factors, on the "vertical variability" typically associated with each state. On the other hand, the adequate number of states to model a certain character depends on the underlying horizontal variability. The possible or optional blank space that may appear between characters should be also modeled by each character HMM. In many cases the adequate number of states may be conditioned by the available amount of training data.

Once an HMM "*topology*" (number of states and structure) has been adopted, the model parameters can be easily trained from images of continuously handwritten text (*without any kind of segmentation*) accompanied by the transcription of these images into the corresponding sequence of characters. This training process is carried out using a well known instance of the EM algorithm called *forward-backward or Baum-Welch re-estimation* [10].

Words are obviously formed by concatenation of characters. In our finite-state modeling framework, for each word, a stochastic finite-state automaton is used to represent the possible concatenations of individual characters to compose this word. As previously discussed, the possible inter-character blank space is modeled by the character-level HMM. In contrast with continuous speech recognition, blank space often (but not always) appears between words. This automaton takes into account this possible blank space, as well as optional character capitalizations.

Sentences are formed by the concatenation of words. This concatenation is modeled by an n-gram model [10], which uses the previous $n-1$ words to predict the next one. N-grams can be easily represented by finite-state deterministic automata. N-grams can be max-likelihood learned from a training (text) corpus, by simply counting relative frequencies of n-word sequences in the corpus [10].

As discussed in section 4, all these finite-state (character, word and sentence) models can be easily *integrated* into a single *global* model on which both equations (5) and (6) are easily solved; that is, given an input sequence of raw feature vectors $\boldsymbol{x}$, a pair $(\hat{c}, \hat{s})$ is obtained for the integrated approach, or only $\hat{s}$ is obtained for the recognition phase of the serial approach. The classification phase of the serial approach is done in accordance with equation (7), using n-gram models or multinomial classifiers.

6 Experiments

The image dataset extracted from survey forms consists of 913 binary images of handwritten phrases scanned at 300 dpi. Each of these images was preprocessed as discussed in section 3. Following results reported in [5], a vertical resolution of 1/20 was adopted. Therefore, each phrase image is represented as a sequence of (3×20)-dimensional feature vectors. The resulting set of sequences was then partitioned into a training set of 676 samples and a test set including the 237 remaining samples.

In order to train the models as described in section 5, a transcription of each training image was written. The resulting transcription set accurately describes all the elements appearing in each handwritten text image, such as (lowercase and uppercase) letters, symbols, abbreviations, spacing between words and characters, crossed-words, etc. It was used to train the character HMMs and the N-gram models for both, the integrated recognition-classification approach and the recognition phase of the serial approach.

Table 1. Basic statistics of the database and its standard partition.

Number of:	Training	Test	Total	Lexicon
phrases	676	237	913	–
characters	64,666	21,533	86,199	80
words	12,287	4,084	16,371	3308

All the 80 characters and symbols appearing in the image corpus were modeled using the same left-to-right HMM topology. After informally testing different values, HMMs were configured with 6 states and 64 Gaussian (diagonal) densities per state. On the other hand, integrated recognition-classification 1-gram and 2-gram models using Witten-Bell back-off smoothing [11, 12] were trained from the transcription set. Also from this transcription set, the recognition and the eight classification 1-gram and 2-gram models of the serial approach were trained. Similarly, a multinomial text classifier was trained from this set, using the smoothing techniques described in [7].

Table 2 shows recognition and classification error rates for the integrated and serial approaches, using different combinations of recognition and classification N-gram models. Also, it includes the classification error rate for the serial approach using a multinomial text classifier.

Because of the difficulty of the task, it can be seen that error rate results are high. In general, both approaches yield similar results, around 34% in recognition and 50% in classification with unigrams (bigrams give slightly worse results); i.e., half of decisions are correct while half are wrong. It is a bit better than the

Table 2. Test-set *recognition* word error rate (WER) and classification error rate are given for integrated and serial approaches. Results are reported using unigram and bigram as language models and as text classifier for serial approach. Classification result using the multinomial text classifier is included.

	Integrated Approach		Serial Approach	
	Recog.(WER)	Classif.(ER)	Recog.(WER)	Classif.(ER)
1-gram	34.4	50.2	34.3	51.1
2-gram	33.6	59.1	32.5	57.0
Multinomial	n/a	n/a	n/a	43.0

52% reported in [6], using the serial approach and a slightly different preprocessing and feature extraction. Apart from these results, it is worth noting that the multinomial classifier achieves a 43% of classification error, which is significantly better than the 50% obtained with unigrams.

An extra classification experiment was carried out using the multinomial text classifier on the correct test sample transcriptions (i.e., without recognition errors), resulting in a 40% of classification error rate. As this result does not differ significantly from the 43.0% obtained with recognized phrases, it can be said that the difficulty of this task does not lie on the recognition phase, but rather on the inadequate classification scheme employed and the insufficient amount of data for training the classification models.

Figure 2 shows three examples of handwritten phrase images along with their integrated approach results.

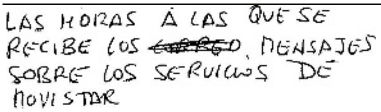
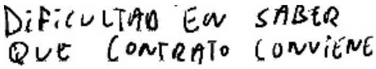
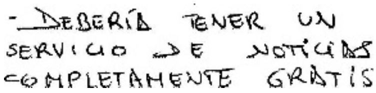
Image	Recognition Result	Classification Result
	**BIEN OTRAS D** LAS QUE SE RECIBE LOS **CORREO** MENSAJES SOBRE LOS SERVICIOS **A P** MOVISTAR	Wrong
	DIFICULTAD EN SABER QUE CONTRATO **CAMBIAR**	Correct
	DEBER-A TENER UN SERVICIO **EN E** NOTICIAS COMPLETAMENTE GRATIS	Correct

Fig. 2. Examples of three handwritten phrases along with their recognition and classification results. The misrecognized words are indicated in underlined bold-face.

The first one produced six word errors and was wrongly classified. The second and third produced one and two word errors, respectively, but both were correctly classified. It is then clear that correctly classified phrases do not imply that a perfect recognition has been achieved. In fact, it is often the case that we get a correctly classified phrase after an imperfect recognition phase.

7 Conclusions

Two approaches have been discussed for a task of spontaneous handwritten text recognition and classification: an integrated one and a serial one. Both are based on Hidden Markov Models and N-grams though, in the case of the serial approach, we have also considered the multinomial Naive Bayes text classifier for the second step. In general, both approaches yield high, similar error rates: around 34% in recognition and 50% in classification with unigrams (43% with the multinomial classifier in the serial approach). This is due to difficulty of the task and, in particular, of its text classification subtask.

References

1. Guillevic, D., Suen, C.Y.: Recognition of legal amounts on bank cheques. Pattern Analysis and Applications **1** (1998) 28–41
2. Bazzi, I., Schwartz, R., Makhoul, J.: An Omnifont Open-Vocabulary OCR System for English and Arabic. IEEE Trans. on PAMI **21** (1999) 495–504
3. González, J., Salvador, I., Toselli, A.H., Juan, A., Vidal, E., Casacuberta, F.: Off-line Recognition of Syntax-Constrained Cursive Handwritten Text. In: Proc. of the S+SSPR 2000, Alicante (Spain) (2000) 143–153
4. Marti, U.V., Bunke, H.: Using a Statistical Language Model to improve the preformance of an HMM-Based Cursive Handwriting Recognition System. Int. Journal of Pattern Recognition and Artificial In telligence **15** (2001) 65–90
5. Toselli, A.H., Juan, A., Keysers, D., González, J., Salvador, I., H. Ney, Vidal, E., Casacuberta, F.: Integrated Handwriting Recognition and Interpretation using Finite-State Models. Int. Journal of Pattern Recognition and Artificial Intelligence **18** (2004) 519–539
6. Toselli, A.H., Juan, A., Vidal, E.: Spontaneous Handwriting Recognition and Classification. In: Proceedings of the 17th International Conference on Pattern Recognition. Volume 1., Cambridge, United Kingdom (2004) 433–436
7. Juan, A., Ney, H.: Reversing and Smoothing the Multinomial Naive Bayes Text Classifier. In: Proc. of the 2nd Int. Workshop on Pattern Recognition in Information Systems (PRIS 2002), Alacant (Spain) (2002) 200–212
8. Pastor, M., Toselli, A., Vidal, E.: Projection profile based algorithm for slant removal. In: International Conference on Image Analysis and Recognition (ICIAR'04). Lecture Notes in Computer Science, Porto, Portugal, Springer-Verlag (2004) 183–190
9. Cavnar, W.B., Trenkle, J.M.: *n*-gram-based text categorization. In: Proc. of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR-94), Las Vegas, Nevada, U.S.A. (1994) 161–175
10. Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press (1998)
11. Katz, S.M.: Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. IEEE Trans. on Acoustics, Speech and Signal Processing **ASSP-35** (1987) 400–401
12. Witten, I.H., Bell, T.C.: The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression. IEEE Trans. on Information Theory **17** (1991)

Combining Global and Local Threshold to Binarize Document of Images

Elise Gabarra[1] and Antoine Tabbone[2]

[1] AlgoTech' Informatique, Technopole Izarbel, Hôtel d' entreprises 1, 64210 Bidart, France
[2] Université Nancy 2, LORIA, Campus Scientifique,
B.P. 239, 54506 Vandoeuvre-lès-Nancy Cedex, France
e.gabarra@algotech.fr, Antoine.Tabbone@loria.fr

Abstract. In this paper, a new approach to binarize grey-level document images is proposed. The method combines a global and a local approaches. First, we provide the edges of the image, and next, from the edges we make a quadtree decomposition of the image. On each area of the image, a local threshold is computed and applied to all the pixels belonging to the region under consideration.

Introduction

Segmentation of gray-level images consists in making a partition of the image into several homogeneous regions. Usually, the different regions distinguish each others according to criteria which permitted to build them[12].

Binarization is a particular case of segmentation. In image processing, binarization's aim consists in making two regions, one made by objects (information) and another one made by the background. In some applications the use of I-level images decreases the computational cost of subsequent processing steps. There is several methods for segmentation[7,12]:

- methods based on grey-level of regions,
- methods based on edge detection,
- methods based on thresholds computed for different regions.

Most of methods are either based on global or local threshold. In global approaches a threshold is calculated and applied to all the pixels of the image[15]. Most of these methods use statistical methods like Bayes classifier or maximum likelihood[4,7,9], moment preservation[14], signal processing (maximization of entropy of the image, minimization of the variance between the object and the background[10]), Hadamard transform[2] and multi-scale histogram separation[13]. For some applications, local approaches[12] are more accurate. Local methods use different thresholds according to the region under consideration. However they often rely on a size parameter which may change for different images or for different locations inside the same image providing in some cases too noisy results.

Binarized documents can be of different qualities. The nature of the support (paper, tracing paper…) and the degradation of documents due to their stocking and lighting conditions and their uses, influence the quality of the documents. That's why a simple

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 371–378, 2005.

threshold won't give systematically good results. The approach we propose here, tries to suit a wide range of documents. For this purpose we combine a global and a local methods. In the next two sections we describe our approach. First we explain how the edges are detected in an image and after we give an algorithm to divide the contour image into a quadtree. Then, experimental results are provided and compared to others methods and finally, we draw conclusions and propose directions for future research.

Edge Detection

Edges research in a numeric image, is one of the most studied problem since the first works in numeric imaging. This is mainly due to the very intuitive nature of edge which appears naturally like the ideal visual clue in the largest cases.

The approach we present here, is based on Canny's edge detection[1] who proposed an edge detector. The detector should satisfy three criteria:

- The edge detector should respond only to edges, and should find all of them. No edge should be missed (Error rate).
- The distance between the edge pixels found by the edge detector and the actual edge should be as small as possible (Localization).
- The edge detector should not identify multiple edge pixels where only a single edge exists (Response).

The maximisation of these criteria leads to the resolution of a differential system whose solution is a filter f which looks like the derivative of a Gaussian function:

$$f(x)=-(x/\sigma^2)\exp(-x^2/2\sigma^2) \tag{1}$$

As the Gaussian is separable, the convolution in two dimensions can be separated into two convolutions in one dimension. An approximation of this filter has been implemented using IIR filter by R. Deriche[5].

The first step in this edge detection consists in making a smoothing of the image to get out the noise which contaminates the image. That's why a Gaussian filter is used, because it's a good compromise between spatial localization and frequential localization. So a two-dimensional Gaussain mask is created to be convolved with the image. The standard deviation of the Gaussian is a parameter set manually by the user.

The second step consists in computing the magnitude at each pixel of the smoothed image. The derivative filter is a one-dimensional mask convolved with the image in the lines' direction and in the columns' direction. Two images are obtained. The magnitude is computed by the combination of the X and the Y components of the gradient and we obtain an image of pixels magnitudes:

$$M(x,y)=\sqrt{G_X(x,y)^2+G_Y(x,y)^2} \tag{2}$$

The third step concerns the "nonmaximum suppression". An edge is detected where the variation of grey-level is significant. The magnitude, which is the variation speed of the intensity for each pixel, will determine if a pixel will be selected or not to belong to the edge.

The magnitude of the pixel gradient must be higher than its neighbours' one, towards the orientation of the pixel. The pixel (*x,y*) is a local maximum if and only if:

$$\begin{cases} M(x,y) \geq M(x_d, y_d) \\ M(x,y) \geq M(x_{d+\pi}, y_{d+\pi}) \end{cases} \tag{3}$$

with (x_d, y_d) and $(x_{d+\pi}, y_{d+\pi})$ the neighbours of (*x,y*), in the gradient direction *d*.

The magnitude is estimated from the gradients of the neighbouring pixels. It's assumed that the gradient is a linear function. Then the gradient can be approximated by a linear interpolation.

In the final step, we apply an threshold method called "hysteresis". It consists in computing a double threshold (a high and a low) and applying it to the image containing only the local maximum points. Every pixel whose grey-level is above the high threshold, is a starting-point and every pixel connected to this starting-point, whose grey-level is above the low threshold, belongs to the edge.

This two thresholds a high threshold are automatically defined from the gradient modulus histogram distribution. The high threshold is set at 90% (respectively at 10% for the low) of pixels on the magnitude histogram. Experimental results have shown that if the high threshold is lowered and the low threshold increased, the filter will be more sensible to the grey-level variations. But it will be also more sensible to the noise.

Quadtrees

At this stage, we have an image of edges. We recall that our aim is to define locally a threshold for each region delimited by a closed contour. So, our first idea was to close the edges and to determine the corresponding areas enclosed by a closed edge. However, after some experimental tests, it appears to be very difficult to describe a textured area with a enclosed contour. For example, the edges detected in Fig 1 for the textured area do not mean anything in our case. So it is difficult to close such edges in order to describe a region.

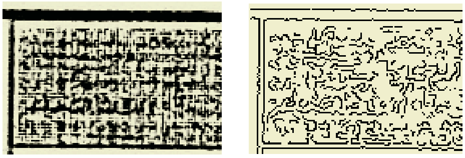


Fig. 1. Example of textured area.

For this reason, we focus our attention to quadtrees methods. In a first time, the whole image is decomposed into four areas called quadrants or nodes. If one of this four quadrants contains information, (that's to say edges) the quadrant is subdivided in its turn, into four quadrants, otherwise, the quadrant is not subdivided and it becomes a terminal node. While the criteria of presence of edges in a quadrants is validate, the recursive decomposition is applied to this quadrant. Another criterion is applied in this decomposition, the size of the region must be greater than a threshold called ε to avoid to small regions containing only few pixels.

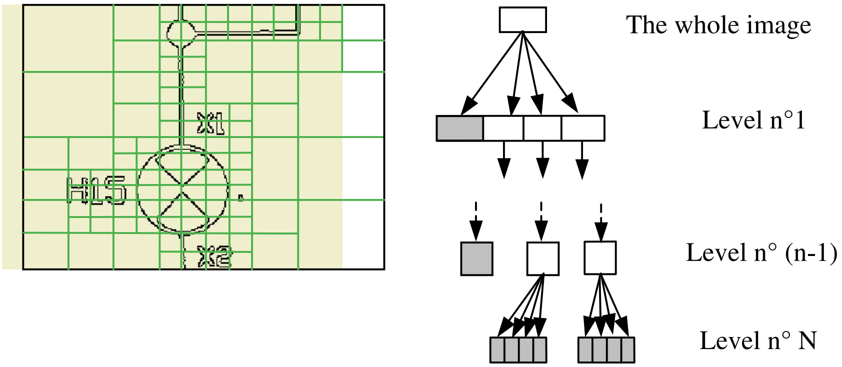


Fig. 2. Example of quadtree result for an edges image.

This recursive decomposition is represented in a tree data structure. At the top level of the tree, level number 0, we found a *quadrant* which represents the whole edges image. At the level number 1, there is the four *quadrants* from the first decomposition and so on. At last, we have a tree with N levels.

Once the decomposition is over, a threshold is computed for each terminal node. We start by the last level (level N in Fig 2) and we proceed all the terminal nodes defined at this level. For those one which contain edge pixels, the threshold is an average of the grey-levels corresponding to the grey-levels of the pixels from the smoothed image.

For the terminal nodes which do not contain any edges, the threshold is the average of the neighbours areas thresholds containing edges, and belonging to the same level. nth level (the same level). For the terminal node of the others levels (from the (n-1)th level down to the level number 0), the threshold is computed as follows : the threshold of each area located in the ith level is computed by making the average of the neighbours' terminal nodes thresholds, which belong to the previous levels (lower levels).This is done recursively until a threshold has been calculated for each terminal node. Fig. 2 illustrated our quadtree decomposition algorithm on a edge image.

Then the whole image is binarized following the local threshold defined locally on each node. If the grey-level of a pixel from the smoothed image is above the threshold of its corresponding area, the pixel belongs to the objects else to the background.

Experimental Results

First, we have compared our results with ones obtained with the best manual threshold it is possible to find globally on the image.

In the first histogram (see Fig.3.b), we see easily two areas (marked by circles), one corresponding to the background (the biggest one) and the other one to objects. But it seems to be difficult to find a global threshold to separate the image into two regions. We have set manually two thresholds (see Figs 5 and 6).

To compare our approach with a global threshold we try to find manually the best threshold to separate the objects from the background.

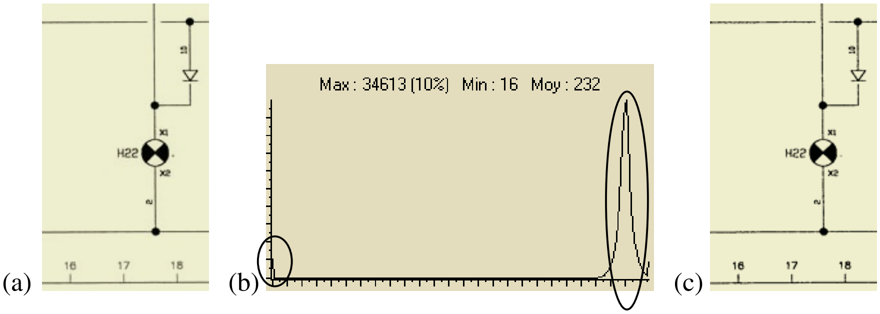


Fig. 3. (a) Test image. (b) Grey-level histogram. (c) Image binarized with our method.

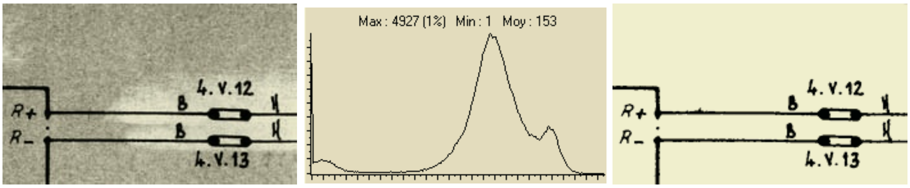
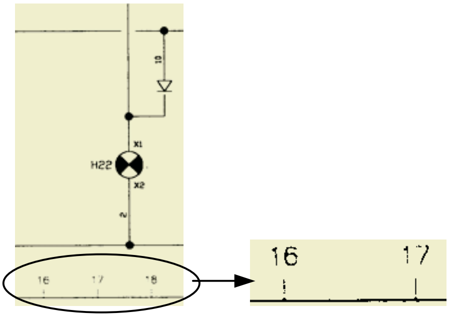


Fig. 4. (a) Test image. (b) Grey-level histogram. (c) Image binarized with our method.



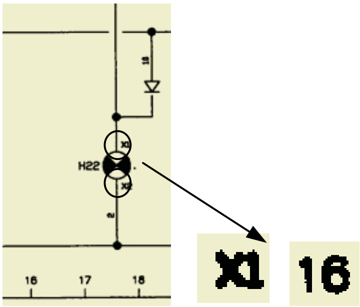


Fig. 5. Binarized image with a threshold=128. **Fig. 6.** Binarized image with a threshold=192.

With a threshold around 128, we loose some information. Less pixels belong to objects than in the result obtained with a threshold of 192. That's why some information are in the background. Although, with the threshold around 192, more pixels are taken to make part of objects. But several pixels should rather belong to the noise.

With the second image, we have also tested two thresholds. With a threshold around 60, the noise is not completely deleted. If we decrease the threshold, the noise disappears but some information disappear too (see Figs 7 and 8). To conclude it is difficult for this kind of image to find a global threshold which provides results similar to the results achieved with our approach.

We have compared our method with three others methods. The first one is proposed by Tabbone and Wendling[13]. It's a multi-scale algorithm based on a statistical test of homogeneity which permits to decide whether a region belongs to the background or not. Stable regions in scale space are used as a model to automatically find a threshold from the intensity histogram. The second one is proposed by Trier and Taxt[15]. The Laplacian's polarity is used to label pixels as being objects or background. Only pixels having a high gradient modulus (called the activity threshold by Trier and Taxt) are considered and zero-marked regions are labelled as objects or background according to the 8-connected neighbours.

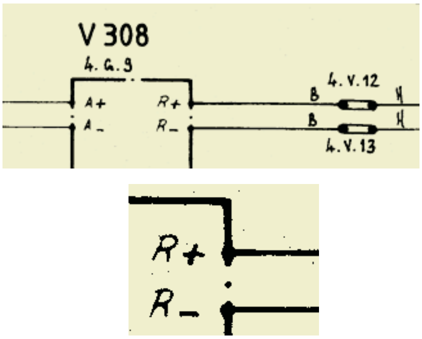


Fig. 7. Binarized image with a threshold=60.

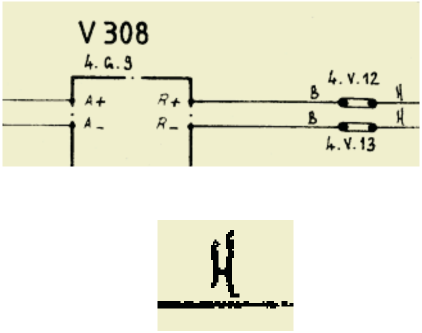


Fig. 8. Binarized image with a threshold=35.

The last method is proposed by Cheng and Chen[3]. It consists in making a fuzzy partition on a 2D histogram of the image. The partition criteria are based on the optimization of the fuzzy entropy.

To compare the performance of the different approaches, we use two quality criteria proposed by Levine and Nazif[8]. This criteria are measures of goodness of the segmentation without knowledge on the underlying objects belonging to the image. These two criteria are:

– A measure of contrast : $$C_I = (1/\#\text{Regions}) \sum_{Rj \in I} | m_{Rb} - m_{Rj} | \qquad (4)$$

Where I is the grey-level image, R_j is the jth segmented region, m_{Rb} is the mean of the region assumed as background after the binarization process *and* m_{Rj} is the mean of the connected components $R_{j.}$ A large value of C_I indicates a high inter-regions contrast.

– A measure of homogeneity : $$H_I = \sum_j \sum_{(x,y) \in Rj} (I_{(x,y)} - m_{Rj})^2. \qquad (5)$$

A low value of the homogeneity means a high intra-regions uniformity.

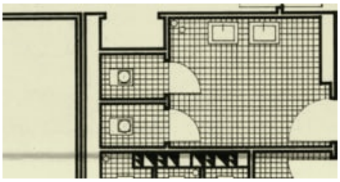


Fig. 9. Test image.

Results presented in Fig. 11 show that the method presented in [13] has the best intra-regions uniformity. However, our method presents a high inter-regions contrast because a threshold is adaptatively defined on each node of the quadtree.

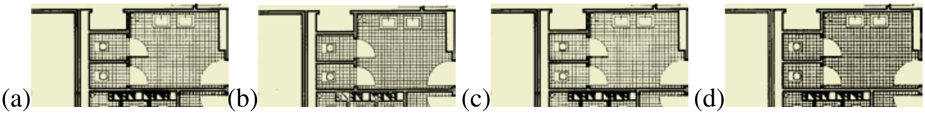


Fig. 10. (a) results with Tabbone and Wendling method[13] (b) results with Trier and Taxt method[15] (c) results with Cheng and Chen method[3] (d) results with our method

	Tabbone and Wendling	Trier and Taxt	Chang and Chen	Our method
Ci	87.03	44.57	85.06	118.48
Hi	756.19	1823.84	866.38	1385.30

Fig. 11. Performances of the four methods tested on the images provided in Fig 10.

Conclusion and Perspectives

The proposed approach is original and provides good results on various kinds of images. The use of local and global information proposed here, provides better results than a single global or local threshold. The binarized symbols in the document are more precisely localised. Further investigations will extend the proposed method in two ways.

The first idea consists in using an adaptative document-smoothing. Instead of applying a classical linear smoothing filter, we should apply a filter which adapts the smoothing process to the contrast, orientation, and spatial size of local image structures.

The second idea concerns the quadtree approach. The threshold ▯ we use to represent the minimum size of an area, could be replaced by statistical tests which would evaluate automatically the uniformity of the region area[6].

References

1. J.F. Canny, 1986. A computational approach to edge detection. IEEE Transaction on PAMI 8(6): 679-698.
2. Chang, F., Liang, K.-H., Tan, T.-M., Hwang, W.-L., 1999. Binarization of documents images using Hadamard multi-resolution analysis. In: Proceedings of 5th International Conference on Document Analysis and Recognition, Bangalore, India September pp. 157-160.
3. Cheng, H.D., Chen, Y.-H., 1999. Fuzzy partition of two-dimensional histogram and its application to thresholding. Pattern recognition 32, 825-843.
4. Cho, S., Haralick, R., Yi, S., 1989. Improvement of Kittler and Illingworth's minimum error thresholding. Pattern Recognation 22, 609-617.
5. R. Deriche, 1987. Using Canny's criteria to derive a recursively implemented optimal edge detector. The International Journal of Computer Vision 1(2):167-187.
6. Gadi T. and Benslimane R., 2000. Fuzzy hierarchical segmentation *(in French)*. Traitement du signal 17(1).
7. Kurita, T., Otsu, N., Abdelmalek, N., 1992. Maximum likelihood thresholding based on population mixture models. Pattern Recognition 10, 1231-1240.
8. Levine, M.D., Nazif, A.M., 1985. Dynamic measurement of computer generated image segmentation. IEEE Transactions on PAMI 7 (2), 155-164.

9. Mardia, K.V., Hainsworth, T.J., 1988. A spatial thresholding method for image segmentation. IEEE Transactions of PAMI 10 (6), 919-926.
10. Otsu, N., 1979. A threshold selection method for grey-level histograms. IEEE Transactions on Systems, Man, and Cybernetics 9, 62-66.
11. Parker J., 1996. Algorithms for Image processing and computer vision. Wiley Editions.
12. Sauvola J., Pietikainen M, 2000. Adaptative document image binarization, Pattern Recognition 33, 225-236.
13. Tabbone S., Wendling L., 2001. Multi-Scale binarization of images. Pattern Recognation Letters 24, 403-411.
14. Tsai, W., 1985. Moment-preserving thresholding: a new approach. Computer Vision, Graphics and Image Processing 29, 377-393.
15. Trier, D., Taxt, T., 1995a. Improvement of “integrated function algorithm” for binarization of document images. Pattern Recognition Letters 16, 277-283.

Extended Bi-gram Features in Text Categorization*

Xian Zhang and Xiaoyan Zhu

Department Of Computer Science and Technology, Tsinghua University,
Beijing 100084, P.R. China
zx97@mails.tsinghua.edu.cn

Abstract. Usually, in traditional text categorization systems based on Vector Space Model, there is no context information in a feature vector, which limited the performance of the system. To make use of more information, it is natural to select bi-gram feature in addition to unigram feature. However, the longer the feature is, the more important the feature selection algorithm is to get good balance in feature space This paper proposed two feature extraction methods which can get better feature balance for document categorization. Experiments show that our extended bi-gram feature improved system performance greatly.

1 Introduction

More and more textual documents are available in internet, which makes it more difficult to manage text data and to retrieve useful information from document contents. Text categorization is an important way to help resolve this problem, which is an increasingly important field and has been extensively studied.

Many statistical classification methods and machine learning techniques have been applied to text categorization in recent years [1]. Most of current text categorization techniques are based on Vector Space Model, in which a document is converted to a high dimensional vector composing a Feature Vector Space. Usually, the components of the feature vector space are isolated words. These systems treat a document as a "Bag of Words" (BOW) instead of a "Sequence of Words", and perform classifications based on some statistical weight of the features [2]. On some well-organized test corpora, such as Reuters news corpus, several systems based on Vector Space Model work well with just very few features [1]. Koller shows that in a hierarchical setting, documents in Reuters corpus can be classified very accurately with just 10 words [6]. McCallum also shows that, using a Naïve Bayesian classifier, many of the main categories in the Reuters corpus can be classified accurately with less than 100 words [7]. However, it is not true for the database collected from web pages, since the documents are in more free style and have much bigger vocabulary set. Chakrabarti reports that, on a selected sample set from Yahoo, the system that works well on Reuters corpus get a poor precision of 32% [3]. Sahami made the same observation in his investigations [5]. In general, IR techniques that work well on standard text corpora often do not on the web information processing.

In the Vector Space Model, the order of the feature dimensions is not important. The original context information between the words is lost while building feature vectors,

* This paper is supported by Natural Science Foundation No.60272019 and No. 60321002

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 379–386, 2005.

which break up the relationships between the words. Since the categorization task is a kind of semantic understanding work, the lack of context information may lead to damage for the system performance. Intuitively, using bi-grams, phrases, as well as n-grams or even sentences, which contain some context information, instead of single words, may improve the characteristic of the feature. But obviously, the long units are large in amount, and have many synonymies, which make worse statistical quality. In addition, the distribution of long units in the documents is badly imbalanced and sparse. Therefore, it is reasonable that long-term features could not make a good classification performance by themselves alone.

In recent years, some researcher suggested that it could improve the classification accuracy by using word pairs or phrases **in addition to** the single word features. A. Aizawa selected appropriate Compound Words as features by the help of Brill Tagger and some predefined rules [4]. In Chade-Meng Tan's experiments, selected unigrams are used as seeds to generate bi-gram features [9]. Their experiments show that properly selected phrases as unigram's patches may outperform the simple unigram feature models, and the point is how to choose the good feature subset from the large amount of original feature space.

Text categorization task is usually a multi-class classification problem and could be solved by dealing with several 2-class problems. Here, the training documents form a positive data set with target class, and the rest is the negative data set. Then there are two types of features, global feature and local feature [8]. Global feature, selected from the whole dataset, generates a universal feature space for all 2-class problems. Moreover, local feature is generated for each category respectively. It is believed that usually local feature is better than global feature. However, global feature takes every class into account, which makes the feature space more balanced, and somewhat more stable than local feature space.

A major difficulty of feature selection is the high dimensionality of the feature space. The unique words that occur in document, which construct the native feature space, can be hundreds of thousands of terms even in a moderated-sized text corpus. When considering bi-grams or phrases, the number will be much larger. This is prohibitively too high for many learning algorithms so that automatically reducing the native space without sacrificing categorization accuracy is required.

This paper made a deep discussion on the weakness of traditional features. Then two kinds of features, CBF and PBF, are devised and tested to remedy the problems. Section 2 describes the feature selection problem in text categorization, mainly on the weakness of IG selection. Section 3 presents two methods that extend the original bi-gram features to CBF and PBF features to re-balance the feature set. Section 4 gives out the experiments and the results. Section 6 summarizes the conclusions.

2 Features

In Vector Space Model, a document is treated as a Bag of Words [2], and the text categorization problem now can be solved by classifying the vectors in feature space made up of words. However, the quality of the features is one of the most important factors. The BOW model is based on the hypothesis that all the features are statistically independent, which is definitely not true in real world. All positional information is lost in the BOW model if the feature space contains only the single words. In this means, bi-gram, which is made up of adjacent unigrams, might be better than unigram, for reserving context information. But there are several weakness for bi-gram. First, the dimensionality of the bi-gram feature is too large and very sparse. In mean time, many bi-gram items are frequently used in only a few articles, so the distribution is very imbalanced. Third, there are more synonyms and lastly, phrases tend to be more noisy because of the redundancy. For all these reasons, bi-grams must be properly selected before being used as a feature.

Information Gain, infogain in short, is a statistical property that is used to measures the worth of an attribute for a classifier. Given the corpus, *S* containing examples from *n* classes, the infogain of *S* is:

$$Infogain(S, A) = Entropy(S) - \sum_{v \in Values(A)} P(v) Entropy(S_v) \tag{1}$$

$$Entropy(S) = \sum_{i=1}^{n} -P(c_i) \log_2 P(c_i) \tag{2}$$

where $P(\cdot)$ should be the probability, whereas the proportion is used in stead of. As mentioned, we have converted the multi-classification task into several 2-class problems. Formula (1) can be rewritten as follows in detail:

$$IG(w) = C + P(w) f[P(c_+ \mid w)] + P(\overline{w}) f[P(c_+ \mid \overline{w})] \tag{3}$$

where $C = Entropy(S)$ and $f(x) = x \log_2 x + (1-x) \log_2 (1-x)$. *C* is a constant for the same corpus. The marks "+" and "-" refer to the positive and negative data, respectively. $P(w)$ is the proportion of the documents where word *w* occurs, while $P(\overline{w})$ is for *w* not occurring, $P(c_j \mid w)$ is for the documents assigned to class c_j in the documents with feature *w*, and $P(c_j \mid \overline{w})$ is for class c_j that do not contain *w*. If the infogain of a feature is zero, the distribution of the feature is all the same in all the categories. That means the feature provides no helpful information for classification.

In most cases, the data in positive set is from just one of the tens of categories, while the negative data is the combination of all the other categories. Therefore, $P(c_+)$ is a very small value. Usually we have $P(c_+) < 0.1$. As our experiment result shows, the Term Frequency of most words is no more than 50 in almost 14000 documents, which leads to a very small $P(w)$. We define *NumOf*(*w*) as the number of documents that contain feature *w*, and *NumOf*($\overline{w}$) for those without *w*. We have:

$$NumOf(w) << NumOf(\overline{w}) \tag{4}$$

Then we can get $P(w) << P(\overline{w})$. That means the second item in formula (3) is much smaller compared with the third. As a result, only the third item is considered in the following discussion.

In practice, $P(c_+ \mid \overline{w})$ and $P(c_+ \mid w)$ could be calculated as follows:

$$P(c_+ \mid w) = \frac{NumOf(wc_+)}{NumOf(w)}, \quad P(c_+ \mid \overline{w}) = \frac{NumOf(\overline{wc_+})}{NumOf(\overline{w})} \tag{5}$$

where wc_+ indicates the documents that contain word *w* and belong to positive set and $\overline{w}c_+$ without word *w*. Similarly, we use $\overline{w}c_-$ and wc_- to indicate the document belong to negative sets with or without word *w*. In such 2-class model, the positive data is all documents about one topic, so it is easy to find common words and phrases. However, the negative class contains many topics, which makes it impossible to find common terms among most of the topics. While selecting features, we want to get features that can represent the topic of the category. Nevertheless, as shown later, it is very difficult to find a feature that can represent the topic of mixed negative set. When considering bi-gram feature with worse statistical properties than unigram, it might be more troublesome to select useful bi-grams.

First think about the features related to the topic of positive set, which should occur frequently in the positive set. Usually, $NumOf(wc_+)$ would be comparable with $NumOf(\overline{w}c_+)$, which can be formulated as:

$$NumOf(\overline{w}c_+) = \alpha NumOf(wc_+) \tag{6}$$

where the parameter α is a positive and not far from 1. Then $P(c_+ \mid \overline{w})$ can be rewritten as:

$$\begin{aligned} P(c_+ \mid \overline{w}) &= \frac{NumOf(\overline{w}c_+) + \dfrac{NumOf(w)}{NumOf(\overline{w})} NumOf(\overline{w}c_+)}{NumOf(w) + NumOf(\overline{w})} \\ &< \frac{NumOf(\overline{w}c_+) + NumOf(wc_+)}{NumOf(All)} = \frac{NumOf(c_+)}{NumOf(All)} = P(c_+) < 0.1 \end{aligned} \tag{7}$$

Consequentially, the infogain of *w* is larger than $\overline{w}$'s.

In contrast, in the case of the features that do not related to positive set, we have:

$$NumOf(\overline{w}c_+) >> NumOf(wc_+) \tag{6'}$$

Then we could get a bigger $P(c_+ \mid \overline{w})$, with similar considering as formula (7), and subsequently a smaller infogain. Usually, the feature with bigger infogain is selected firstly. That is why most of the selected features are related to positive dataset. Tan's experiments also confirmed this [9]. In fact, it is reasonable that the selected features are imbalanced since the imbalance of the training data, specially, for bi-gram features. Our approach is to resolve is problem as described in the next section.

3 Extended Bi-gram Features

In this paper, two kinds of feature, CBF and PBF, are achieved by proposed algorithms to rebalance the feature set.

3.1 Combined Bi-gram Feature

We notice that the bi-gram feature selected above is related to the category topic too much. In other words, they are too "local" for the category. To fix this problem, we must make the features more "global". Suppose there are k categories, $c_1,\ldots, c_k$ in the corpus S. We can get k bi-gram feature sets, labeled $B_i\,(i = 1,2,...,k)$, which is extracted between c_i and the other k-1 categories. As analyzed above, B_i is too much related to the category c_i, while little to the other categories. Intuitively, if we combine all feature sets, except B_i, after filtering, to make a feature complement for c_i, we can get more "global" features. The algorithm for selecting **Combined Bi-gram Feature(CBF)** is shown in the following:

Step 1. For all $c_i\,(i$=1,…$k)$, extract bi-gram feature set B_i={(b_{ij}, ig_{ij})| ig_{ij} = *Infogain*(S, b_{ij}) } between c_i and all the other documents.
Step 2. For all $i \neq t$, combine all (b_{ij}, ig_{ij}) pairs in B_i into a temporary set T. If b_{ij} has existed in T, just add ig_{ij} to current infogain value.
Step 3. All the bi-grams in T whose accumulative infogain are bigger than an infogain gate, *ig_min,* are selected as CBF feature of category c_t.
where *ig_min* is determined by experiments.

3.2 Pairwise Bi-gram Feature

Another important reason of the imbalance in bi-gram features is the difficulty to find common terms in the documents including different topics. Therefore, we can extract the bi-grams not between the target topic and the others, but between only two categories in a pairwise way. So we can get $\frac{1}{2}k(k-1)$ feature subsets. The last thing is how to combine these feature subsets to construct the final feature space. The algorithm for selecting **Pairwise Bi-gram Feature(PBF)** is as following, where c_t is the target topic:

Step 1. For all $i \neq t$, build bi-gram feature B_{it} ={(b_{ij}, ig_{ij})| ig_{ij} = *Infogain*(S, b_{ij}) } set between categories c_i and c_t.
Step 2. For all $i \neq t$, add all triples (b_{ij}, ig_{ij}, n_j) to a temporary set T, where (b_{ij}, ig_{ij}) is from B_{it} and n_j is the number of times that bi-gram b_{ij} has occured. If b_{ij} has existed in T, just add ig_{ij} to current infogain value, and increase n_j by 1.
Step 3. Those bi-grams in T that make inequation (8) true are selected as the final **PBF** feature of category c_t.

$$\frac{ig}{t}\alpha^{t-1} \geq ig\_gate \quad (\alpha > 1) \tag{8}$$

The parameter α can be achieved by following simple iteration:

Step 1. Initialize α to a positive number a little bigger than 1. For example, set α=1.1.

Step 2. Sort the triples in temporary set T according to value $\frac{ig}{t}\alpha^{t-1}$ in descending order. If all bi-grams with n_j=k-1 is in front of the first position where the bi-gram with n_j=k-3 occurs, output α and stop the iteration.
Step 3. Slightly increase α. For example, set $\alpha = \alpha$+0.1. Goto step 2.
where the *ig_gate* is determined by experiment.

4 Experiments and Results

In the experiments, the dataset Yahoo! Science hierarchy, called “Yahoo! Science” dataset, is used as [7, 9]. The corpus is originally contains 14,869 documents in 30 top-level categories. We focus our attention on the 10 largest categories and all other documents are moved to make an "Others" category. The standard performance measures for Text Categorization are **recall, precision** and **F1-measure**. Recall is the percentage of total documents for the given topic that are correctly classified, while precision is the percentage of predicted documents for the given topic that are correctly classified [1]. F1-measure is a helpful measure for evaluating the effectiveness of classifiers, which is give by:

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{9}$$

Experiments are carried out on each categoriy by simple unigram features, original selected bi-gram features, and our two types of extended bi-gram features: **CBF** and **PBF**, respectively. We get 8190 CBFs and 8252 PBFs on average of all categories. Table 1 lists some of the selected PBFs of Agriculture category. We can see that not only phrases about Agriculture are selected, but also phrases that have little relationship to Agriculture, e.g., “space+flight” and “nasa+gov”, which may represent the characteristics of negative document set, are in the feature space.

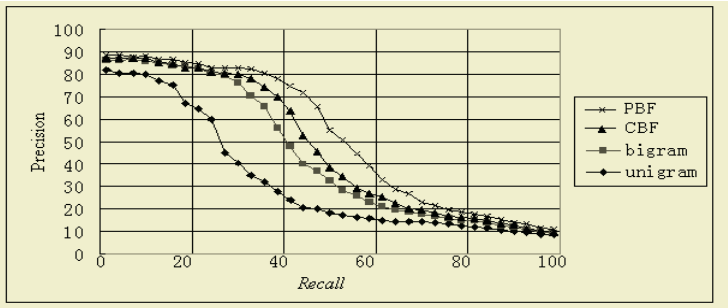


Fig. 1. The Recall-Precision curves of the experiments on Agriculture category

Table 1. Top 20 PBF triples in temporary set T in descending order of value ${}^{ig}\!/_{t}\ \alpha^{t-1}$ in the experiment of Agriculture Category

n_j	*Infogain*	**PBF feature**	n_j	*Infogain*	**PBF feature**
9	0.246064	natural+resource	9	0.078927	agricultural+research
9	0.244462	cooperative+extension	8	0.090603	educational+programs
9	0.202362	department+agriculture	3	0.182231	nasa+gov
9	0.171706	agriculture+natural	8	0.089125	remington+electronic
9	0.108873	sustainable+agriculture	8	0.089125	marti+Remington
9	0.108034	animal+science	8	0.080219	department+animal
8	0.128333	university+nebraska	8	0.076953	research+extension
9	0.09885	college+agriculture	8	0.074426	field+crops
8	0.122425	electronic+publications	8	0.073025	extension+service
8	0.118107	nebraska+lincoln	4	0.1313	space+flight

Table 2. The maximum F1-measure scores in all categories using four kind of features

Category	**Unigram**	**Bi-gram**	**CBF Feature**	**PBF Feature**
Agriculture	0.4364	0.5302	0.5726	**0.6173**
Astronomy	0.6926	0.7550	0.7796	**0.7820**
Biology	0.5396	0.6034	0.6202	**0.6279**
Computer Science	0.4814	0.5946	0.6120	**0.6506**
Earth Sciences	0.6440	0.6976	**0.7172**	0.7165
Engineering	0.6958	0.7186	0.7330	**0.7335**
Mathematics	0.4990	0.6064	0.6250	**0.6565**
Physics	0.5264	0.6072	0.6480	**0.6641**
Space	0.6212	0.6416	0.6446	**0.6657**
Zoology	0.8326	0.8356	**0.8654**	0.8612
on average	0.5989	0.6609	**0.6817**	**0.6975**

Fig. 1. shows the Recall-Precision curves of the classification results for Agriculture. The classification result using PBF is absolutely better than the others. Table 2 lists all the experiment results. It also proves that in most cases, PBF feature gets the best result. This is reasonable because the PBF features are extracted between every two topics, which makes the selected features more precisely related with category topic. On average, the F1-measure scores using our extended feature sets improve by 13.8% and 16.5% over the baseline score of the unigram feature sets, and our results are higher than simple bi-gram features by 3.15% and 5.54%.

5 Conclusion

This paper presents two methods to extract more balanced bi-gram features, CBF and PBF, for overcoming the imbalance in training data sets. Experimental results show that the approaches using our extended bi-gram features outperform those using traditional unigram features and simple bi-gram features on the database collected from web pages. In future, more detailed experiments should be done to check whether these feature could be used in common data bases.

References

1. Yiming Yang, Xin Liu. A re_examination of text categorization methods Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99, pp 42--49), 1999
2. Lewis, D. Representation and Learning in Information Retrieval Technical Report UM-CS-1991-1993. Department of Computer Science, University of Massachusetts, Amherst, MA
3. Chakrabarti S., Dom B., and Indyk P. Enhanced hypertext categorization using hyperlinks. In SIGMOD 1998, Seattle, Washington. 1998
4. A. Aizawa. Linguistic techniques to improve the performance of automatic text categorization. In Proceedings 6th NLP Pac. Rim Symp. NLPRS-01
5. Sahami M. Using Machine Learning to Improve Information Access. PhD Thesis (1998). Stanford University, Computer Science Department

6. Koller D. and Sahami M. Hierarchically Classifying Documents Using Very Few Words. In ICML-97: Proceedings of the Fourteenth International Conference on Machine Learning, pp.170-178, San Francisco, CA: Morgan Kaufmann. 1997
7. McCallum A. and Nigam K. A Comparison of Event Models for Naive Bayes Text Classification. AAAI-98 Workshop on "Learning for Text Categorization". 1998
8. Dumais S T, Platt J, Heckerman D, et al. Inductive learning algorithms and representations for text categorization. Technical report, Microsoft Research. 1998
9. Chade-meng Tan, Yuan-fang Wang, Chan-do Lee. The Use of Bi-grams to Enhance Text Categorization Journal Information Processing & Management, 2002

Combining Fuzzy Clustering and Morphological Methods for Old Documents Recovery

João R. Caldas Pinto[1], Lourenço Bandeira[1], João M.C. Sousa[1], and Pedro Pina[2]

[1] IDMEC, Instituto Superior Técnico
Av. Rovisco Pais, 1049-001 Lisboa Portugal
{jcpinto,lpcbandeira,j.sousa}@dem.ist.utl.pt
[2] CVRM / Geo-Systems Centre, Instituto Superior Técnico
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
ppina@alfa.ist.utl.pt

Abstract. In this paper we tackle the specific problem of old documents recovery. Spots, print through, underlines and others ageing features are undesirable not only because they harm the visual appearance of the document, but also because they affect future Optical Character Recognition (OCR). This paper proposes a new method integrating fuzzy clustering of color properties of original images and mathematical morphology. We will show that this technique leads to higher quality of the recovered images and, at the same time, it delivers cleaned binary text for OCR applications. The proposed method was applied to books of XIX Century, which were cleaned in a very effective way.

1 Introduction

The universal availability and the on-demand supply of digital duplicates of large mounts of written text are two inevitable paths of the future. To achieve this goal books have to be first digitalized. Due to the rarity of most of them this operation should be carried out only once, so we obtain high resolution color images. These large dimension images will be the future source data. The first natural possibility of the libraries is to make available in the Internet reduced versions of these images. However, not only the size is a problem, but also the visual appearance of the books can be of poor quality due to the aging process. Spots due to humidity, marks resulting from the ink that goes through the paper, rubber stamps, strokes of pen and underlines are features that, in general, are not desirable. By the other hand, it is an ultimate goal of the libraries to obtain ASCII versions of the books, what means to perform optical character recognition (OCR). However, if this is a well-established process for images of texts written with modern fonts and on clean paper, it is still a challenging problem for degraded old documents. Thus, all the contributions to improve quality of the images are important to the success of this process.

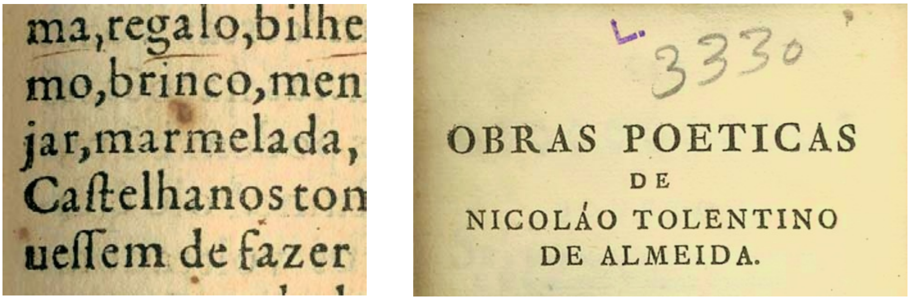
In this paper we address the problem of documents image enhancement. In general, an old book contains a short number of colors, typically only two: one for the ink and another for the background, the paper color. However, with the ageing process colors change, the ink becomes clearer and the background yellowed, and other colors emerge, some due to natural causes, like humidity or print through, and other by human actions, such as pencil handwritten notes, rubber stamps or underlines.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 387–394, 2005.

This is illustrated in Fig. 1. This suggests that images can be segmented by colors using clustering algorithms, like fuzzy clustering [1-3]. On the other hand, from the geometric point of view, characters are quite distinct from the other elements present in a page. This fact can be exploited through a mathematical morphology approach [4, 5].

This paper proposes a combination of fuzzy clustering of original color images followed by a mathematical morphology step for removing residual geometrical artifacts. Applying our method we achieve two goals: an improved document appearance and a high quality binary image for further application of OCR, using for example the Fine-Reader Engine [6].

This paper is organized as follows. Section 2 describes the fuzzy clustering techniques used in this paper, and their application to document segmentation. Section 3 describes the application of mathematical morphology to the same problem. The proposed recovering algorithm is presented in Section 4, where the obtained results are described and discussed. Finally, Section 5 presents the conclusions and possible future work.



(a) natural ageing processes (b) human manipulation

Fig. 1. Example of a printed page with common problems

2 Fuzzy Clustering

Fuzzy clustering (FC) in the Cartesian product space is applied to partition the data into subsets. Cluster analysis classifies objects according to similarities amongst them. In image analysis, clustering finds relationships between the system properties (colors in this paper).

Let $\{x_1, ..., x_N\}$ be a set of N data objects where $x_k \in \mathbb{R}^n$. The set of data objects can then be represented as a $N\times n$ data matrix X. The fuzzy clustering algorithms determine a fuzzy partition of X into C clusters. Let $U = [\mu_{ik}] \in [0,1]^{N\times C}$ denote a fuzzy partition matrix of X. Often, the cluster prototypes are points in the cluster space, i.e. $v_i \in \mathbb{R}^n$. The elements $\mu_{ik} \in [0,1]^{N\times C}$ of U represent the membership of data object x_k in cluster i. Let V be a vector of cluster prototypes (centers) to be determined, defined by $V = [v_1, v_2, \ldots, v_C]$.

Many clustering algorithms are available for solving for *U* and *V* iteratively. The fuzzy c-means is quite well known and revealed to present good results [1]. This algorithm does not determine directly the optimal number of clusters. This paper uses simple heuristics to determine the correct number of clusters, in order to reduce the number of colors classifying the different samples of text images. The fuzzy c-means algorithm searches for an optimal fuzzy partition *U* and for the prototype matrix of cluster means *V*. In other words,

$$\left(X,C\right)\xrightarrow{clustering}\left(U,V\right) \tag{1}$$

The optimization minimizes the following objective function,

$$J\left(X,U,V\right)=\sum_{i=1}^{C}\sum_{k=1}^{N}\left(\mu_{ik}\right)^{\alpha}d_{ik}^{2} \tag{2}$$

where α is a weighting parameter. The function d_{ik} is the distance of a data point x_k to the cluster prototype v_k : $d_{ik}^2=\left(x_k-v_i\right)^T\left(x_k-v_i\right)$. The fuzzy c-means algorithm can be described as follows.

Given the data *X*, choose the number of clusters $1<K<N$, the fuzziness parameter $m>1$ and the termination criterion $\varepsilon>0$. Initialize $U^{(0)}$ (e.g. random).

Repeat for $l=1,2,\ldots$

Step 1: **Compute cluster means:**

$$v_i^{(l)}=\frac{\sum_{k=1}^{N}\left(\mu_{ik}^{l-1}\right)^m x_k}{\sum_{k=1}^{N}\left(\mu_{ik}^{l-1}\right)^m},\quad 1\le i\le K \tag{3}$$

Step 2: **Compute distances** for $1\le i\le K\ \ 1\le k\le N$.

$$d_{ik}^2=\left(x_k-v_i^{(l)}\right)^T\left(x_k-v_i^{(l)}\right) \tag{4}$$

Step 3: **Update partition matrix** for $1\le i\le K\ \ 1\le k\le N$.

$$\mu_{ik}^{(l)}=\frac{1}{\sum_{j=1}^{K}\left(d_{ik}/d_{jk}\right)^{2/(m-1)}} \tag{5}$$

until $\left\|U^{(l)}-U^{(l-1)}\right\|<\varepsilon$.

3 Mathematical Morphology

In the present context, Mathematical Morphology (MM) is applied to remove handwritten underlines before the OCR phase and is based on a previous study [7]. Two main steps constitute this phase: the first one consists of reinforcing the text set whose segmentation sometimes produces irregular and broken characters, while in the second one the underlines are suppressed.

Step 1: **Text characters reinforcement**

This objective is achieved by firstly applying a closing (φ) with the structuring element B_1 of size λ to the initial binary image X, resulting from the fuzzy clustering phase, in order to reinforce the characters:

$$Y_1 = \varphi^{\lambda B_1}(X) \tag{6}$$

The filtering of unwanted structures of smaller dimension than text characters (defined by the dimension of the structuring element B_2) is also a necessary operation to carry out, and is obtained by an erosion (ε) – reconstruction (R) sequence. The result is given by set Y_2:

$$Y_2 = R_{Y_1}[\varepsilon^{\lambda B_2}(Y_1)] \tag{7}$$

Step 2: **Handwritten underlines removal**

The handwritten underlines are marked by applying directional openings $\gamma^{\lambda l}(Y_2)$ with a segment l as structuring element in the horizontal direction (0 degrees). Only horizontal or sub-horizontal structures can resist, totally or partially, to this transform [8]. The partial directional reconstruction (DR) of the remaining regions permits to recover the underlines: the geodesic dilation uses a directional structuring element and is applied till the idempotence is reached. The set difference with the image Y_2 permits to filter out the handwritten underlines. This sequence is summed up in the following equation:

$$Y_3 = Y_2 / DR_{Y_2}[\gamma^{\lambda l(o^{\circ})}(Y_2)] \tag{8}$$

In order to recover the regions of the characters suppressed by the elimination of the handwritten underlines (in these regions there exists a superimposition between letters and underlines), a dilation $\delta^{\lambda l}$ in the vertical direction is applied. It gives the possibility of recuperating partially these common regions without reconstructing again the handwritten structures:

$$Y_4 = \delta^{\lambda l(90^{\circ})}(Y_3) \tag{9}$$

The resulting set constitutes now the binary image to be introduced in the OCR system.

4 Results

Integrating both methods we can take advantage of the most positive aspects of both approaches. The integration method uses the binarization of the output image of the FC step as the input image of the MM, as it can be seen in Fig. 2. In this way, the input image of the MM is the result of a very efficient segmentation process. The result of this MM step is a high quality cleaned binary image that can also be used for OCR applications [9].

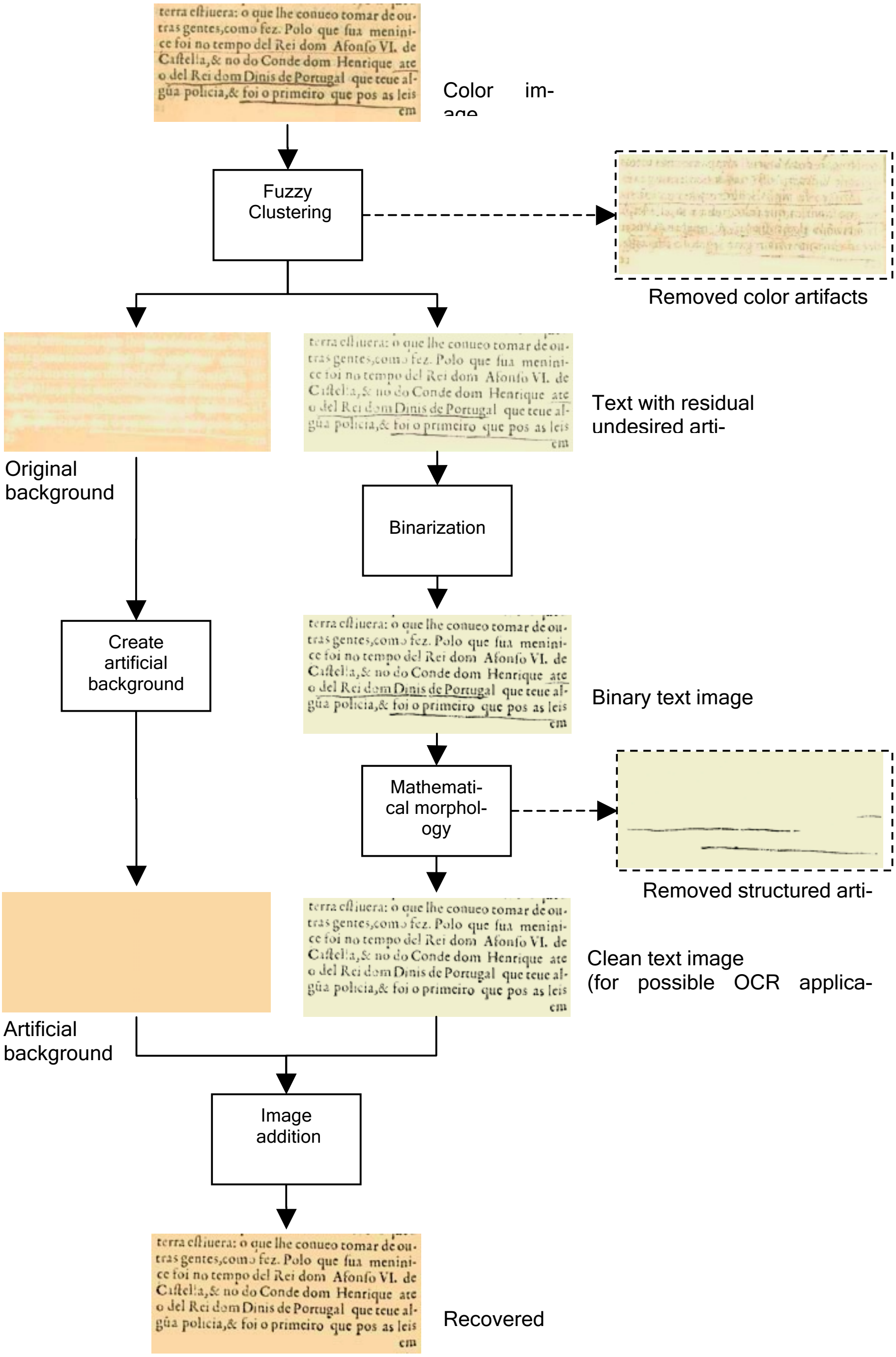
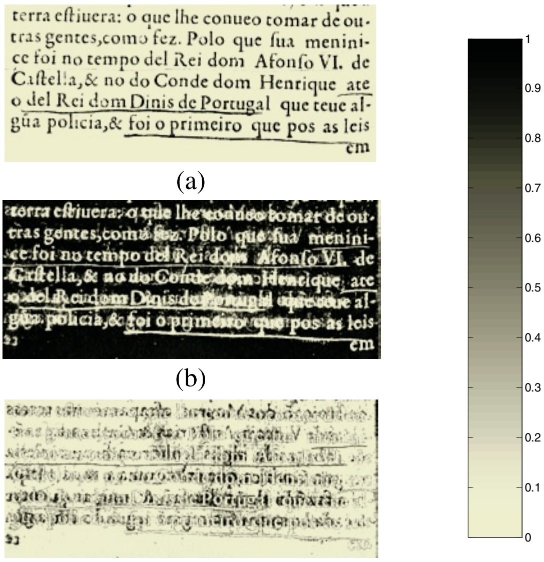


Fig. 2. Fluxogram of the proposed recovering algorithm

In order to test the proposed technique, fuzzy clustering (FC) allied to mathematical morphology (MM), several experiments were conducted. The performance of each step was tuned and evaluated by visually inspecting the preprocessed word images.

With the FC approach, three clusters have been used, because three distinct regions can be easily identified with the human eye: the background, the characters and the image defects, such as underlines, humidity spots, see-through letters, etc. The segmentation results obtained with the FC are presented in Fig. 3.



(a)

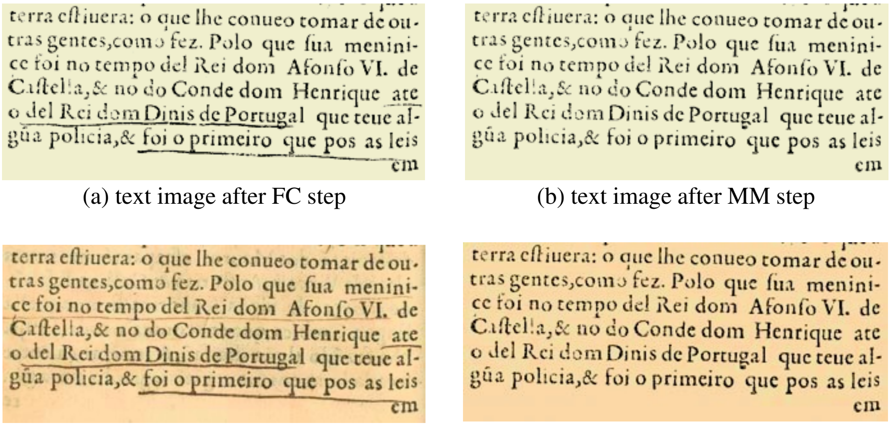
(b)

(c) (d)

Fig. 3. Segmentation results: (a) text cluster, (b) background cluster, (c) cluster with undesired color artifacts and (d) gray scale representing membership to cluster *i*

Each cluster is classified through the analysis of the average and variance of every pixel that belong to more than 85% to that cluster. The characters (represented by the darker cluster) and the background (represented by the lighter cluster) can be easily identified by the cluster's average. With the background's variance, an artificial background was reproduced using a Gaussian distribution with the known parameters. To achieve this purpose we suggest a solution based on the fact that histograms corresponding to the RGB components of a homogeneous region approximately follow a Gaussian distribution with standard deviation smaller then 10 [10]. Results confirm the correctness of this approach (see Fig. 2 and Fig. 4). Note that if we wish to keep some other image features like handwritten notes we only need to keep a higher number of resulting clusters.

An overall inspection of the images obtained by the FC step (Fig. 4(a)) and the MM step (Fig. 4(b)) permits to conclude that both methods are very satisfactory in removing features, since all of the undesired color/structured image defects are removed. The FC step produces normally clean background images with a good visual aspect. However, in some situations, there exists an over-filtering by suppressing some pixels of some words. Not only the MM step helps to remove undesired residual structural elements but also corrects some of these "damages" introduced in the characters by the FC. Comparing all four images in Fig. 4. it can be seen that the integration algorithm takes advantage of the most positive aspects of both methods.



(a) text image after FC step

(b) text image after MM step

(c) original color image

(d) recovered color image

Fig. 4. Images from different steps of the algorithm

5 Conclusions

Recovering the visual appearance of degraded old documents is an important issue because it is a real concern of the Libraries to make them available by digital means, particularly through Internet access. Simultaneously, clean documents are an important contribution to a higher OCR performance when applied to old documents, a problem now being tackled by several methods but still facing great challenges.

We proposed a novel solution based on the combination of fuzzy clustering of the original images and a mathematical morphology step for removing residual geometrical artifacts. From the obtained results, we can conclude that this approach achieves very good outcomes. In addition, a high quality binary image is produced. These images can afterwards be used as inputs to OCR algorithms contributing to better performances. We emphasize that this method leads in several occasions to better quality binary images than FineReader and its only drawback is to be slightly more time consuming.

As future work, some improvements can still be done in order to make this software available to the librarians, which is the ultimate goal of this project. In particular, some manually parameterization should be automated. Finally, the proposed algorithm will be extended to multicolored pages and characters.

Acknowledgements

This work was partly supported by: the "Programa de Financiamento Plurianual de Unidades de I&D (POCTI), do Quadro Comunitário de Apoio III"; the FCT project POSI/SRI/41201/2001; "Programa do FSE-UE, PRODEP III, Quadro Comunitário de Apoio III"; and program FEDER.We also wish to express our thanks to the Portuguese Bibioteca Nacional for their continuous support, which made possible this work.

References

1. Bezdek J. C.: Pattern Recognition with Fuzzy Objective Function. Plenum Press, New York (1981)
2. Buse R., Liu Z. Q., and Bezdek J.: Word recognition using fuzzy logic. IEEE Transactions on Fuzzy Systems, 10(1) February (2001) 65–76
3. Driankov D., Hellendoorn H., and Reinfrank M.: An Introduction to Fuzzy Control. Springer, Berlin (1993)
4. Serra J.: Image Analysis and Mathematical Morphology. Academic Press, London (1982)
5. Soille P.: Morphological Image Analysis. 2nd edition. Springer, Berlin (2003)
6. ABBYY FineReader Homepage, http://www.abbyy.com, ABBYY Software House
7. Caldas Pinto J. R., Pina P., Bandeira L., Pimentel L., Ramalho M., Underline Removal on Old Documents, Lectures Notes in Computer Science, LNCS 3211, Springer, (2004) 226-234
8. Soille P., Talbot H.: Directional Morphological Filtering. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(11) (2001) 1313-1329
9. Ribeiro C.S., Gil J.M., Caldas Pinto J.R, Sousa J.M.: Ancient document recognition using fuzzy methods. In: Proceedings of the 4th international Workshop on Pattern Recognition in Information Systems, Porto, Portugal (2004) 98-107
10. Caldas Pinto J.R., Marcolino A., Ramalho M., Clustering Algorithm for Colour Segmentation, SIARP'00 - V Ibero-American Symposium On Pattern Recognition, (2000) 611-617

Part V

Bioinformatics

A New Algorithm for Pattern Optimization in Protein-Protein Interaction Extraction System*

Yu Hao[1], Xiaoyan Zhu[1], and Ming Li[2]

[1]State Key Laboratory of Intelligent Technology and System,
Department of Computer Science and Technology,
Tsinghua University, Beijing, China
haoyu@s1000e.cs.tsinghua.edu.cn,
zxy-dcs@tsinghua.edu.cn
[2] School of Computer Science,University of Waterloo,
Waterloo, Ont. N2L 3G1, Canada
mli@uwaterloo.ca

Abstract. In pattern matching based Protein-Protein Interaction Extraction systems, patterns generated manually or automatically exist erroneous and redundancy, which greatly affect the system's performance. In this paper, a MDL-based pattern optimizing algorithm is proposed to filter out the bad patterns and redundancy. Experiments show that our algorithm is effective in improving the system's performance while greatly cutting down the number of patterns. It also has excellent generalizability which is important in implementing practical systems.

1 Introduction

In recent years, many accomplishments have been made in building biological literature data mining systems, most of which are *Protein-Protein Interaction* systems. Correspondently, algorithms and methods are developed to ensure the systems to work accurately, efficiently and robustly [1][2][3]. Correspondently, algorithms and methods are developed to ensure the systems to work accurately, efficiently and robustly, which are based on three main approaches, i.e. Bayesian[5], NLP[4], and Pattern Matching[1]. Although the three approaches have been developed well respectively for many years, once integrated into biological literature mining, there exist problems and limitations with the consideration of the complexity of biological sentences' grammatical structure [3]. Pattern matching based systems, are simple and robust ones, which map the part-of-speech sequences to the structural information slots according to the predefined patterns and matching rules. There are two kinds of approaches in generating these patterns, manually drafting and automatically extracting. Ono proposed a pattern matching based system which used manually coded rules of simple world and part-of-speech patterns to extract special kind of protein interactions from abstracts, and achieved high recall and precision rate for yeast(recall =

* This paper is Supported by Natural Science Foundation No.60272019 and No. 60321002.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 397–404, 2005.

82.5%, precision = 94.3%)[1]. In GENIES, more complicated patterns with syntactic and semantic constraints are drafted, but the recall rate is relatively lower[6]. Manually draft patterns are not practical for users to build a perfect pattern set due to the increasing amount of literature and various writing styles. Huang, however automatically discovers patterns from literature texts by implementing the sequence alignment methods based on *Dynamic Programming* algorithm, whose performances are better than hand coded pattern matching approaches and shows good robustness according to the increasing amount of literature data[7].

Of all the above approaches, pattern optimization is an important issue, because there exist erroneous and redundant patterns, especially when the pattern number is large. Ono evaluated each pattern respectively by the criteria of their precision, which is effective in get rid of some bad patterns, but can't declined the redundancy, such that patterns related may affect each other and lower the system's performance. Huang, in his systems, imposes some grammatical rules to get rid of those ill-formed patterns, but the rules which are drafted manually, may incur errors and delete good patterns, thus is less effective. Further more, in above all approaches, the patterns which are evaluated and optimized on the training set, usually has lower performance in the testing set without considering the issue of generalizability.

In this paper, we propose a new pattern evaluation and optimization algorithm based on MDL (*Minimum Description Length*) principle to evaluate each pattern according to the performance of the whole pattern set. Our approach is all data driven, and the MDL principle ensures that it has better generalizability. Experiment shows that the optimized pattern number is greatly reduced by our algorithm, while the system's precision improves a lot without much lost in the recall rate on both training and test set.

The paper is arranged as follows: in section 2, the structure of our system is introduced, proposed MDL-based optimizing algorithm are described in section3, three experiments which tests the effectiveness of our algorithm are presented in section 4, and section 5 is the discussion and future work.

2 System Overview

Our protein-protein interaction extraction system is divided into 3 phases, the Data Preparation Phase, the Pattern Generation Phase, and the Interaction Extraction Phase.

The Data Preparation phase first converts the input sentence into tagged sequences for pattern generation and interaction extraction.

Then the Pattern Generation Phase mines the tagged sequences in the corpus and extracts patterns. All patterns are tag sequences, where each tag is called a *component*. Except for the PTN, all the tag has a word set which contains the words that can be instantiated. For example, a pattern {PTN VBZ IN PTN: *; binds, associates; to, with;* }, the word set of tag VBZ is { binds, associates }, while tag IN is {with}[7]. The acquired patterns are then evaluated and optimized by the proposed MDL-base algorithm, and the resulting ones are stored in the Pattern Database for the use of Interaction Extraction Phase, which extracts interactions based on *Dynamic Programming* matching algorithm at last.

3 Pattern Evaluation and Optimization

In pattern matching based systems, there should be a pattern set $P = \{p_1, p_2, \dots, p_m\}$ consisting of candidate patterns p_i, which are extracted from literatures manually or automatically. But there is no guarantee that all of them are correct and works without any redundancy. If a pattern produces too many errors, it is a 'bad' pattern and should be got rid of or modified. On the other hand if it can be replaced by other patterns without affecting system performance, it is a redundant pattern and should be deleted or be merged into other ones to reduce pattern's complexity. Because simpler patterns mean better generalizability and worse accuracy, the optimizing task becomes a trade-off problem between the above two.

Rissanen[8], proposed the Minimum description length(MDL) principle as an powerful tool to solve the trade-off problem between the generalizability and the accuracy. The MDL principle doesn't need the analytical form of the risk function, but only count the bits instead. It is in simple form and effective to our problem.

3.1 MDL Principle

The MDL principle states that given some data D, the best model (or theory) M_{mdl} in the set $\mathcal{M}$ of all models consistent with the data is the one that minimizes the sum of

- The length in bits of description of the model, and
- The length in bits of the description of the data with the aid of the model.

It can be express in a formula as below:

$$M_{mdl} = \arg\min_{M} K(M) + K(D|M)\,, \tag{1}$$

where $K(\bullet)$ is the Kolmogorov Complexity[9].

The MDL principle is one of general guidance to the solving of problems of model selection and parameter regression. It is philosophical succinct in nature but is no magic. Paul M. B. Vitányi and Ming Li [9] has proved that the *Exception-Based* MDL can be vindicated and reduced to the MDL principle of (1) under the circumstances of "supervised learning". It is effective to our task of pattern set optimization, and the optimal pattern set P^* is obtained as follows

$$P^* = \arg\min_{P} K(P) + \log_2 d(I\,,\, I^*)\,, \tag{2}$$

where I and I^* are the extracted and expected interaction sequences respectively, $d(I, I^*)$ is the number of differences between I and I^*, .

3.2 Pattern Set Optimization

The pattern set is optimized through MDL as shown in formula (2) , which consists of 2 components, $K(P)$ and $d(I\,,\, I^*)$. For the convenience, we assume $DL(P) = K(P) + \log_2 d(I\,,\, I^*)$ as the description length system, then the optimization task becomes to minimize the $DL(P)$ by the adjustment of parameter P. In order to get $DL(P)$, $K(P)$ and $d(I\,,\, I^*)$ should be calculated respectively, where $d(I\,,\, I^*)$ is the amount of errors occurred when extracting interactions using the pattern set P, which consists of the number of wrong interactions and missed ones, shown as follows:

$$\begin{aligned} d(I, I^*) = {} & N_{\text{wrong}} + N_{\text{miss}} = (N_{\text{extracted}} - N_{\text{correct}}) + (N_{\text{expected}} - N_{\text{correct}}) \\ & = N_{\text{extracted}} + N_{\text{expected}} - 2\, N_{\text{correct}}, \end{aligned} \tag{3}$$

where N_{wrong} is the number of the wrong interactions extracted, N_{miss} is that missed (should be interacted, but not), N_{expected} is the number of interactions expected to be extracted, $N_{\text{extracted}}$ is the total number of interactions actually extracted including correct and wrong, and $N_{\text{corrected}}$ is the number of interactions correctly extracted.

Since $K(P)$ is the Kolmogorov complexity of pattern set P and is non-computable, it can be approximated by the code length of the pattern set $P = \{p_1, p_2, \ldots, p_m\}$, and $p_i = m_i^1\, m_i^2 \ldots m_i^{c(p_i)}$, where $c(p_i)$ is the number of components of pattern p_i such that:

$$K(P) \approx \sum_{i=1}^{m} \sum_{j} | m_i^j |, \text{ where } | m_i^j | = \begin{cases} 1, & m_i^j = PTN \\ \gamma / c(m_i^j), & others \end{cases}. \tag{4}$$

From (4), it is shown that if a component's word set includes more word, the pattern is more generalized to match the sentences, and is much simpler. $| m_i^j |$ decreases when $c(m_i^j)$ increases, such that m_i^j contribute less to the complexity of P, and $K(P)$ is accordingly decreased.

Once the $DL(P)$ is calculated, we can optimize the pattern set P by minimizing the $DL(P)$ taking the parameter of P. Since the modification of P is ranged over the pattern set space $\mathcal{P}$, the searching space of the optimization methods is very large considering the infinite variation of the pattern components and word sets. For the simplicity and efficiency, we take the 'superfluous then condense' strategy to guide the optimization process, which is shown as follow in two:

1. Generate as many candidate patterns as possible, such that the initial pattern set covers all the appropriate patterns in the system. This can be attained by get rid of all the restrictions imposed in the pattern generation phase.
2. Try to delete patterns in the pattern set by minimizing $DL(P)$, such that the pattern remained are all the most competitive 'good' patterns.

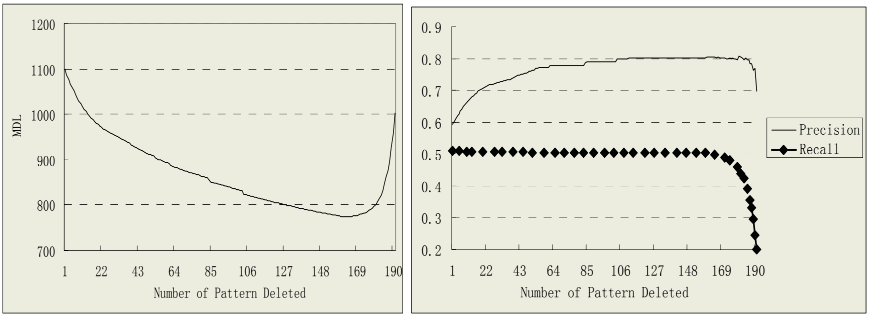
Our minimizing algorithm is guided by the Steepest Gradient Descent strategy which is one of local optimization searching algorithms. In our algorithm, the worst pattern which incurs most increase of the description length of system ($DL(P)$) is deleted in each iteration, until there is no deletion of a pattern can lower the $DL(P)$, when $DL(P)$ reach the minimal value at pattern set P^*, then the optimal pattern set P^* is the one that best fit our system. The whole algorithm is shown as follows:

```
Input: Initial pattern set $P^0$, which contains as many patterns as possible.
Output: optimal pattern set $P^*$
1. $P = P^0$
2. while $P$ is not empty do
    1) Calculate $DL(P)$
    2) MaxDL = -MAXINTEGER
    3) For each $p_i$ in $P$ do
        a) $P^1 = P - \{ p_i \}$
        b) $\Delta DL = DL(P) - DL(P^1)$
        c) if $\Delta DL >$ MaxDL then MaxDL= $\Delta DL$, $i^* = i$
        4) if $\Delta DL<0$ then go to 3 for output
        5) $P = P - \{p_i^*\}$
    3. output $P^*=P$
```

In our approach, a pattern is evaluated and decided whether to be deleted according to its effects to the whole pattern sets *P*, rather than evaluating the pattern individually in its precision and recall rate. Even a pattern that may have good properties, it is deleted if it has redundancies with other patterns in the pattern set and increase the system's description length. So, our algorithm is more effective and P^* is not be the best pattern set that minimize the interaction error but has better generalizability and enhance the system's performance.

4 Experiments

The corpus consists of 963 sentences obtained from internet biomedical papers. Of all these sentences, 1435 interactions are labeled manually. Totally 191 patterns are generated as the initial pattern set without imposing any grammatical rules or optimization algorithm. By implementing our proposed optimization algorithm, each pattern is evaluated and deleted until the optimal point attains. In order to follow the full track of variation of the system description length, precision and recall rate versus the number of pattern deleted, we keep deleting patterns even after the optimal pattern set is obtained, until the pattern set becomes empty. Our experiment is in 3 parts: firstly, the initial pattern set is optimized from all 963 sentences and the description length, precision and recall rate are shown in Figure1. Secondly, the extraction result is compared with that of the rule-based approach. At last, Cross-Validation method is applied to test the generalizability nature of the algorithm.



(a) (b)

Fig. 1. (a) ystem description length variation curve versus the number of patterns deleted. (b) Precision and recall variation curve versus the number of patterns deleted

4.1 Pattern Set Optimization

From figure 1 (a), the minimum description length is obtained at the deleted pattern number of 162, which means there are 30 patterns left in the optimal pattern set. The bottom of the curve (near the optimal point) is flatter than the beginning and ending part because there exist many trivial patterns in the set which make no effect on the extraction process; while the beginning part is steeper, which represents the deletion of bad patterns which produce erroneous interactions. The ending part is the steepest, during which some ‘huge’ patterns that are matched with many sentences to extract most interactions are deleted. For example, pattern {PTN VB IN PTN * ;interact associate ;with ;* ;} is matched to extract 88 interactions, 77 of which are correct. So, most of the errors are inducted by some bad patterns, while very few ‘huge’ patterns takes most of the work of matching. In Ono’s approach, it’s easy to delete those bad ones, but only to preserve the optimal ones seems difficult, where our algorithm can justly do. That also explains why the recall rate curve in figure 1(b) drops dramatically right after the optimal pattern number reached. In figure 1(b), the precision reaches its peak at the optimal point of 162 and drops gently as the precision is more determined by the property of ‘huge’ patterns left in the pattern set.

4.2 Effectiveness Compared with the Rule-Based Approach

Our approach is effective in optimizing the pattern set. According to 4.1, most of the erroneous and trivial patterns are deleted, only those good and ‘huge’ patterns remained in the set. Although manually drafted rules can filter the patterns, it is always hard to find the optimal rules which only preserve those good ones. We compared our approach with the rule-based one in [7], some rules of which are shown as follows:

1. If a pattern has neither verb tag nor noun tag, reject it.
2. If the last tag of a pattern is IN or TO, reject it.
3. If the left neighborhood of a CC tag is not equal the right neighborhood of the tag in a pattern, reject the pattern.

The result is shown in Table 1.

Table 1. System performance of the MDL-based approach and rule-based approach

	Pattern Number	*Training Set*		*Test Set(Cross Validation)*	
		Precision	Recall	Precision	Recall
Original	192	59.2%	50.9%	51.5%	50.0%
Rule-based	65	61.4%	42.3%	-	-
MDL-From Rule-based	12	78.3%	41.7%	-	-
MDL-From Original	30	80.4%	50.1%	73.4%	44.7%
ERM	134	-		64.9%	46.0%

In training set part of table 1, the rule-based approach reduced the pattern number from 192 to 65, while the MDL-based algorithm to 30. The precision of the rule-based approach improves 2.2% from the original while the recall rate declines 8.6%. MDL-based approach improves the precision rate by 21.2%, while the recall rate declines only by 0.8%. It’s clearly seen that the MDL-based approach has much better performance than the rule-base one using even less number of patterns.

The manually drafted rules abruptly declined some good patterns, and allowed many erroneous patterns though they may in good form. So the precision of the rule-based system almost remained the same while the recall rate drops a lot. Following are examples of some good patterns that are deleted by the rule-based approach and otherwise preserved by MDL-based approach:

{IN PTN VBN IN PTN, that ;* ;conjugated interacted ;with ;* ;}
{PTN CC VB IN PTN, * ;and ;associates interacts ;with ;* ;}
{PTN VB CC IN PTN, * ;interact interacts ;and and/or ;with ;* ;}
{PTN VB IN PTN IN, * ;associates interacts ;with ;* ;in ;}

What's more, we further optimize the resulting pattern set of rule-based approach using MDL-based algorithm, and the pattern number is reduced from 65 to only 12, while the precision improves by 16.9% with the recall rate almost at the same. In this test, our methods deletes the erroneous patterns left in the result set of rule-based approach, only good patterns remained.

Our approach is completely data-driven, and is much more effective than rule-based ones, because man can make mistakes while data never.

4.3 Testing Generalizability

The MDL-based algorithm has a good nature in generalizability which is hard to be evaluated with the limited number of 963 sentences and 1436 interactions. Here we implements the strategy of Cross-Validation, and calculate the average performance of all folds. In our experiment 10 folds of sentences are selected circularly, in which the first 70% sentences of each folder are used for optimizing the pattern set, while the last 30% for testing. We also implemented the well known Empirical Risk Minimize (ERM) algorithm to optimize the pattern set as a baseline for comparison. The average system performance is shown as follows in Test set part of Table 1.

From the table, the MDL-based approach reduces the pattern number from 192 to 30, while the precision increases by 11.9% and the recall rate reduced only by 5.3%. It means that the optimized pattern set has better performance in test sets than the original one with so little number of patterns, thus explained the generalizability property of MDL-based algorithm. The ERM-based approach is also examined, where our MDL-based approach improves 8.5% in precision while only decreases 1.3% in recall. The average number of ERM-based optimal pattern set is 134, which is much larger than that of the MDL-based one. In fact the ERM-based optimal pattern set contains too many trivial patterns which are in very complicated form and can only matched to less than 1 sentences in the training set. Those are ill patterns which cause many errors and should be deleted. But the ERM-based algorithm can't get rid of them, while the MDL-based algorithm can, thus it has better generalizability than the ERM-based algorithm.

5 Discussion

The MDL-based optimizing algorithm effectively optimizes the pattern set in greatly improving the precision with only a slight lost of recall rate. It also shows better property of generalizability on the test set. It's a good solution for obtaining optimal patterns, especially automatic generated ones, in protein-protein interaction extraction systems.

However, we only optimize the pattern set by getting rid of the erroneous and redundant ones, thus the recall rate can't be improved during this process. Algorithms of further modifications to the exist patterns is under research, which will not only improve the precision and the recall rate as well. Obviously, the pattern set space has infinite points which make the optimizing process a laborious work. But no matter what searching algorithms implemented, the approach is always based on the framework of MDL principle.

References

1. Ono,T., Hishigaki,H., Tanigami,A., and Takagi,T. Automated extraction of information on protein-protein interactions from the biological literature. Bioinformatics, 17(2), 155–161, 2001
2. Marcotte,EM, Xenarios,I., and Eisenberg,D. Mining literature for protein-protein interactions. Bioinformatics, 17(4), 359–363, 2000
3. Hirschman,L., Park,JC, Tsujii,J, Wong,L., Wu,C.H. (2002) Accomplishments and challenges in literature data mining for biology. Bioinformatics, 18:1553--1561, December 2002
4. Pustejovsky,J, Castano,J, Zhang,J, Kotecki,M, and Cochran,B (2002) Robust relational parsing over biomedical literature: extracting inhibit relations. In Proceedings of the seventh Pacific Symposium on Biocomputing (PSB 2002), pp. 362-373.
5. Ray,S. and Craven,M. (2001) Representing sentence structure in hidden markov models for information extraction. In Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-2001), Morgan Kaufmann, pp. 1273-1279.
6. Friedman,C., Kra,P., Yu,H., Krauthammer,M., and Rzhetsky,A. (2001) Genies: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics, 17 suppl. 1:S74–82.
7. M.L. Huang, X.Y. Zhu, Y. Hao, D.G. Payan, K. Qu, M. Li. Discovering patterns to extract protein-protein interactions from full texts. Bioinformatics. Accepted June, 2004.
8. Yao,D., Wang,J., Lu,Y., Noble,N., Sun,H., Zhu,X., Lin,N., Payan,D.G., Li,M., Qu,K. (2004) Pathway-Finder: paving the way towards automatic pathway extraction. In Yi-Ping Phoebe Chen, ed., Bioinformatics 2004: Proceedings of the 2nd Asia-Pacific Bioinformatics Conference (APBC.)
9. J. Rissanen. Modelling by shortest data description. Automatica, 14:465-471, 1978
10. Paul M. B. Vitányi and Ming Li Minimum Description Length Induction,Bayesianism, and Kolmogorov Complexity IEEE transactions on information theory, vol. 46, no. 2, march 2000

Curvature Based Clustering for DNA Microarray Data Analysis

Emil Saucan[1] and Eli Appleboim[2]

[1] Department of Mathematics, Technion, Haifa and Software Engineering
Department, Ort Braude College, Karmiel, Israel
semil@tx.technion.ac.il

[2] Electric Engineering Department and Department of Mathematics, Technion,
Haifa, Israel
eliap@ee.technion.ac.il

Abstract. Clustering is a technique extensively employed for the analysis, classification and annotation of DNA microarrays. In particular clustering based upon the classical combinatorial curvature is widely applied. We introduce a new clustering method for vertex-weighted networks, method which is based upon a generalization of the combinatorial curvature. The new measure is of a geometric nature and represents the metric curvature of the network, perceived as a finite metric space. The metric in question is natural one, being induced by the weights. We apply our method to publicly available yeast and human lymphoma data. We believe this method provides a much more delicate, graduate method of clustering then the other methods which do not undertake to ascertain all the relevant data. We compare our results with other works. Our implementation is based upon *Trixy* (as available at http://tagc.univ-mrs.fr/bioinformatics/trixy.html), with some appropriate modifications to befit the new method.

Background

Clustering number techniques are extensively employed in the analysis of gene clustering (see [RH]) and in various applications in biomathematics and chemistry (see [WF]). Of particular importance is the *clustering number*, also known as the *core clustering coefficient* or the *combinatorial curvature*, which is defined as follows:

Definition 1. *Let $G = (V, E)$ be a graph, and let $v \in V$. The combinatorial curvature of the vertex v is defined as:*

$$Curv(v) = \frac{|\Delta(v)|}{\rho(v) \cdot (\rho(v) - 1)/2} \tag{1}$$

where $|\Delta(v)|$ represents the number of triangles with apex at v, and $\rho(v)$ is the degree of the vertex v.

Remark 1. Since $|\Delta(v)| \leq \rho(v) \cdot (\rho(v) - 1)/2$, and $Curv(v) \in (0, 1]$, for all $n \geq 2$, $Curv(v)$ is not defined if $\rho(v) = 1$, i.e. if v is a *boundary vertex*.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 405–412, 2005.

But, however simple and intuitive this invariant may be, it fails to take into account the weights of the respective vertices. However, in many practical implementations weights are extremely important (even crucial), such as molecular weight and, more recently taken into consideration, lengths of genes (see [DM]); electric charge; valence of chemical connection; etc.

While representing a generalization of the classical combinatorial curvature, the new curvature we introduce takes into account the weights of any vertex v and its adjacent vertices. This is achieved by introducing a natural *metric* on the vertex-weighted graph, thus rendering G as a *discrete metric space*, and computing its *(metric) curvature.* Thus we give the combinatorial curvature a geometric extension, which allows for a better estimate of local structure and structural proprieties of the graph, since it takes weights into account. Moreover, this provides us with a more sensitive tool, since distances and curvatures take values within a much wider range of *real* numbers – as opposed to the few *rational* values taken by the combinatorial curvature.

Preliminaries

The Metric

We begin by defining the metric on the graph $G = (V, E)$, induced by the weights μ_i associated to the respective vertices $v_i \in V$:

Definition 2. *Let $G = (V, E, \mu)$ be a vertex-weighted graph. Define the metric d by:*

$$d(v,w) = \begin{cases} \frac{|\mu(v)|+|\mu(w)|}{|\mu(v)\mu(w)|} & v \neq w\ ,\ \mu(v), \mu(w) \neq 0; \\ 1 & v \neq w\ ,\ \mu(v) = 0 \text{ or } \mu(w) = 0; \\ 0 & v = w\ . \end{cases} \tag{2}$$

It is straightforward that d satisfies the requirements of a metric on G. (See [Bl].)

Remark 2. In our informational context is natural to choose *natural* (i.e. *positive, integer*) weights, representing the gene lengths, or the atomic (molecular) mass, etc. As such, in this study we restrict ourselves to weights $\mu \in \mathbf{N}$.

Remark 3. If $\mu(v) = 0$, then we are reduced to the combinatorial case, thus it is natural to define $d(v, w) = 1,$ for all $w \sim v$, which represents the usual distance between adjacent vertices for combinatorial graphs.

We begin by taking note of some immediate proprieties of d (remember that $\mu(v) \in \mathbf{N},$ for all $v \in G$):

1. $d(u, v) \in [0, 2]$; for all $u, v \in G$.
2. d(u,v) = 2 iff $\mu(u) = \mu(v) = 1$.
3. If both weights $\mu(v_i), \mu(v_j)$ associated to the vertices of the edge $e = (v_i, v_j)$ are very large, then $d(v_i, v_j)$ approaches 0. (This fits the intuitive fact that "heavy" nodes are "more important".) If, however, only one is large in comparison to the other, e.g. $\mu_i = \mu(v_i)$ then $d(v_i, v_j)$ tends to $1/\mu_i$.

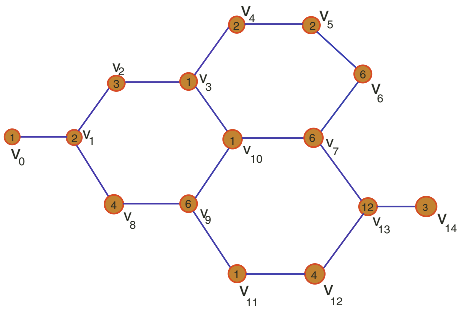


Fig. 1. Distances in a Vertex-Weighted Graph – An example of a vertex weighted graph and the induced distances.

Example 1. Consider the vertex-weighted graph G depicted in Figure 1. Then, to wit, $\mu(v_3) = 1$, $\mu(v_{10}) = 1$, thus by Formula (2) we have $d(v_3, v_{10}) = 2$. Also $d(v_2, v_3) = 4/3$, $d(v_1, v_2) = 5/6$, $d(v_{13}, v_{14}) = 5/12$.

Metric Curvature

The main geometric invariant of a space one wants to investigate is its *curvature*. While classical in the context of smooth surfaces in $\mathbf{R}^3$, the all-important *Gauss curvature* is not natural for discrete metric spaces, as we do have to consider. However, there is a way to adapt the Gauss curvature to discrete metric spaces, which we present below.
We first introduce the classical version of the curvature formula that we want to employ:

Definition 3. *Let (M,d) be a metric space and let $c : I = [0,1] \xrightarrow{\sim} M$ be a homeomorphism, and let $p, q, r \in c(I)$, $q, r \neq p$. Denote by $\widehat{qr}$ the arc of $c(I)$ between q and r, and by qr segment from q to r.*

Then c has Haantjes Curvature $\kappa_H(p)$ *at the point p iff:*

$$\kappa_H^2(p) = 24 \lim_{q,r \to p} \frac{l(\widehat{qr}) - d(q,r)}{\big(l(\widehat{qr})\big)^3} \ ;$$

where "$l(\widehat{qr})$" denotes the length (given by the intrinsic metric induced by d) of $\widehat{qr}$.

Remark 4. (1) While defined for all *rectifiable* curves in any metric space, the Haantjes curvature is defined in fact in a natural, geometric way. Indeed, for the case of smooth curves in the plane ($c \in \mathcal{C}(I, \mathbf{R}^2)$), the Haantjes curvature coincides with the standard curvature of plane curves. (For a short reminder on curvature and the proof of this fact, see [Bl], [BM].)

(2) The 24 factor in the definition above comes from a forth order element in the Taylor-MacLaurin approximation of the curvature.

Adapting this definition to the case of the vertex weighted graphs such that it will represent – as intended – a generalization of combinatorial curvature, i.e. restricting ourselves exclusively to triangles with apex at v, yields the following definition:

Definition 4. *Let $G = (V, E, \mu)$ be as before, let d be the metric on G defined above, and let $v \in V$. Let $\pi = v_1 v v_2$ be a path with v as an internal vertex. We define the curvature of π at vertex v as being:*

$$\kappa'_{H,\pi}(\triangle v_1 v v_2)(v) = \begin{cases} \sqrt{24\frac{|d(v_1,v)+d(v,v_2)-d(v_1,v_2)|}{\left(d(v_1,v)+d(v,v_2)\right)^3}} & e = (v_1, v_2) \in E; \\ 0 & e = (v_1, v_2) \notin E. \end{cases} \tag{3}$$

Then the modified *Haantjes curvature $\kappa'_H(v)$ of G at v is defined to be the arithmetic mean off the curvatures of all the triangles with apex v:*

$$\kappa'_H(p) = \frac{\sum_{\triangle v_1 v v_2} \kappa'_H(\triangle v_1 v v_2)}{\rho(v)(\rho(v)-1)/2} \tag{4}$$

Suppose Δ is a triangle with vertices i, j, k having weights n, m, m respectively. Direct calculation of the triangle curvature at vertex j gives

$$\kappa'_{H,\Delta}(j) = \sqrt{\frac{48m^2(np)^3}{(mn+mp+2np)^3}} \tag{5}$$

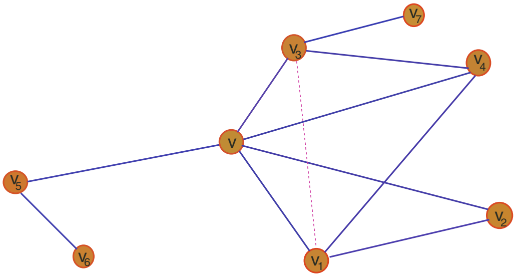


Fig. 2. Haantjes Curvature: Triangles with apex at v and the definition of the modified Haantjes Curvature $\kappa'_H(v)$ of G at v.

Remark 5. (1)In this variation of the definition the curvature at v is computed as the *mean of the curvatures* off all the triangles with apex at v, so in a sense the curvature at each point depends on the curvatures at the points in $v_i \sim v$.
(2) Notice that if we put zero weight at all vertices of the graph then we have constant curvature $= \sqrt{3}$ at all vertices.

Experimental Results

We have implemented the curvature based clustering suggested herein and the results obtained are the subject of this section. Implementation was done by modifying the open source code *Trixy* available from `http://tagc.univ-mrs.fr/bioinformatics/trixy.html`. Modifications where mainly computational so that the computed curvature would be the metric curvature suggested herein, as well as the original computed (combinatorial curvature) as suggested in [RH]. Since inherently from its definition the combinatorial curvature is bounded by 1 we normalized the metric computed curvature in an adaptive fashion with respect to the maximal curvature obtained through the computation.

Clustering was done according to a curvature threshold T_{cur}. Experiments where done on the yeast gene microarrays data available from the website `http://rana.lbl.gov/EisenData.htm`. Building the graphs out of the core data was done according to the method suggested in [RH] using the correlation thresholds employed in the cited work. We do not address this part of the machinery nor did we change this part of *Trixy*.

The weights that are assigned to the vertices should express data that is relevant for the analysis and clustering that we wish to achieve. The data we used here is based on gene length and was extracted from [DM], where gene length was shown to be significant for various functional aspects. The weights where put to the accurate proximity of order yet not necessarily precise to the last digit (e.g. 10,000 instead of say 10,083). Particular emphasize was given to the variance of weights so that it will vary between 100 and 10,000 as in [DM] this variation is claimed to be of extreme importance.

The following two figures show the results of the clustering as processed on a part of the yeast gene expression graph shown in Figure 3.

In Figure $4(a), (b)$ two clustering methods are shown where curvature threshold was set to 0.6, and in Figure $5(a), (b)$ the same methods where applied with curvature threshold 0.7. In both cases the correlation threshold was set as: $T_{cor} = 0.85$.

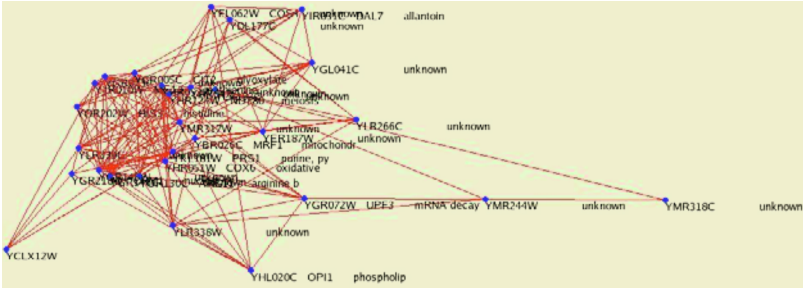
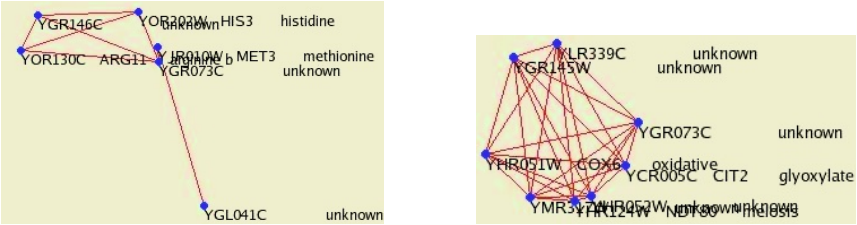
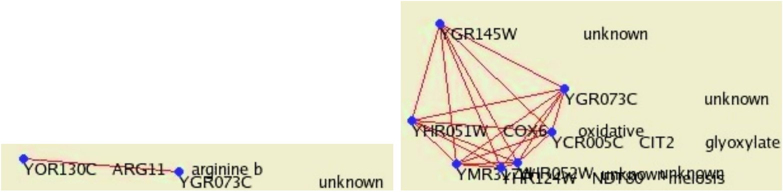


Fig. 3. Part of the Yeast Gene Expression Graph: Part of the yeast gene expression graph with correlation threshold $T_{cor} = 0.85$, as it appears in a snapshot of the main window of the augumented *Trixy*. Both clustering methods depicted in the following two figures were processed upon this graph.



(a) (b)

Fig. 4. Combinatorial (a) and Metric Curvature (b) based Clusterings: The results of the clustering as processed on a part of the yeast gene expression, for $T_{cur} = 0.6$ and correlation threshold $T_{cor} = 0.85$.



(a) (b)

Fig. 5. Combinatorial (a) and Metric Curvature (b) based Clusterings: The results of the clustering as processed on a part of the yeast gene expression, for $T_{cur} = 0.7$ and correlation threshold $T_{cor} = 0.85$.

As can be seen in both cases, metric curvature is more sensitive to minute changes, therefore it preserves more clusters than the combinatorial one.

Discussion and Conclusions

In light of the results above, recall Equation 5, where the curvature at a given vertex of a triangle with given weights at its vertices was calculated. In the following table we give a sketchy analysis for the various possibilities for this weights distribution and their resulting curvatures. This analysis sheds yet additional light on the obtained clustering results.

This results can be partially inferred from the following facts as well as from the table above.

1. Recall that for un-weighted graphs or for graph that are uniformly weighted the metric curvature exactly depicts the combinatorial curvature suggested in [RH]. The results obtained by using metric curvature suggest that yet the graph is further weighted in a meaningful manner, and the curvature measure takes these weights into account the achieved clustering performs better.
2. From the definition of Haantjes curvature we see that the spectrum of possible results is practically a continuum, contrarily to the combinatorial curvature for which only rational number can be realized as valid curvature.

Table 1. Analysis of weights distribution and the resulting curvatures

m, n, p	κ_H
$m \sim n \sim p$	$\kappa_H \sim O(m)$
$m << n \sim p$	$\kappa_H \sim O(1)$
$p << n \sim m$	$\kappa_H \sim O\left(\sqrt{\frac{1}{m}}\right)$
$n << m << p$	$\kappa_H \sim O(1)$
$n << p << m$	$\kappa_H \sim O\left(\sqrt{\frac{1}{m}}\right)$
$p << n << m$	$\kappa_H \sim O\left(\sqrt{\frac{1}{m}}\right)$

Future Research

As noted, both the original combinatorial curvature and the metric curvature as defined in the present work, are defined by considering triangles in graphs. However, as noted in [ESBB], triangles are sparse in generic graphs. Therefore, a more realistic and pliable metric curvature would be able to take into consideration general cycles in graphs, not only triangles. Such a method was devised by the authors in [SA]. Further experiments with this generalized metric Haantjes curvature are planned.

In addition further experiments on larger data sets and their subsequent statistical analysis are currently in process.

Acknowledgements

The authors would like to think Inna Geller, Shirli Kupfberger and Yuval Mironi for their dedicated and skilful programming. They also would like to thank Peter Soreanu for his invaluable collegial help all along this project.

References

[B-DSY] Ben-Dor, A., Shamir, R., and Yakhini, Z. Clustering Gene Expression Patterns. Journal of Computational Biology **6(3/4)** (1999) 281-297.

[Bl] Blumenthal, L.M. Distance Geometry – Theory and Applications. Claredon Oxford (1953)

[BM] Blumenthal, L.M. and Menger, K.: Studies in Geometry. Freeman and Co. San Francisco (1970)

[DM] Duret, L. and Mouchiroud, D.: Expression pattern and, surprisingly, gene length shape codon usage *Caenorharbditis, Drosophila* and *Arabidopsis*. Proc. Nat. Acad. Sci. USA **96** (1997) 4482–4487

[EM] Eckmann, J.-P. and Moses, E.: Curvature of co-links uncovers hidden thematic layers in the World Wide Web. PNAS **99:** (2002) 175–181

[ESBB] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA **95** (1998) **95** 14863–14828

[FJVBO] Farkas, I.J., Jeong, H., Vicsek, T., Barabasi, A.-L., Oltvai, Z.N.: The topology of the transcription regulatory network in the yeast, S. cerevisiae. Physica A (2004) accepted

[HH] Hu, X. and Han, J.: Discovering Clusters from Large Scale-Free Network Graph. preprint (2004)

[HS] Hartuv, E. and Shamir, R.: A Clustering Algorithm based on Graph Connectivity. Information Processing Letters **76** (2000) 175–181

[RH] Rougemont, J. and Hingamp, P.: DNA microarray data and contextual analysis of correlation graphs. BMC Bioinformatics **4** (2003) 15

[SA] Saucan, E. and Appleboim, E.: Can One See the Shape of a Network? – Geometric Viewpoint of Information Flow preprint (2004)

[WF] Wagner, A. and Fell, D.A.: The small world inside large metabolic networks. Proc. R. Soc. Lond. B. **268** (2001) 1803–1810

Support Vector Machines for HIV-1 Protease Cleavage Site Prediction

Loris Nanni and Alessandra Lumini

DEIS, IEIIT – CNR, Università di Bologna
Viale Risorgimento 2, 40136 Bologna, Italy
lnanni@deis.unibo.it

Abstract. Recently, several works have approached the HIV-1 protease specificity problem by applying a number of classifier creation and combination methods, from the field of machine learning. In this work we propose a hierarchical classifier (HC) architecture. Moreover, we show that radial basis function-support vector machines may obtain a lower error rate than linear support vector machines, if a step of feature selection and a step of feature transformation is performed. The error rate decreases from 9.1% using linear support vector machines to 6.85% using the new hierarchical classifier.

1 Introduction

HIV-1 protease [1] is an enzyme in the AIDS virus that is essential to its replication. The chemical action of the protease takes place at a localized active site on its surface. HIV-1 protease inhibitor drugs are small molecules that bind to the active site in HIV-1 protease and stay there, so that the normal functioning of the enzyme is prevented.

Understanding HIV-1 protease cleavage site specificity is very desirable, because efficiently cleaved substrates are also excellent templates for synthesis of tightly binding chemically modified inhibitors.

The standard paradigm for protease-peptide interaction is the "lock and key" model where a sequence of amino acids fits as a key to the active site in the protease, which in the HIV-1 protease case is estimated to be eight residues long.

For these particular problems only one class (the uncleaved category) is shift invariant, the other class is not. Shift invariance means that the category remains unchanged if we shift left or right the pattern of one position. For instance, the peptide KVFGRCELAAAMKRHGLDN is not cleaved by HIV-1 protease, which means that all the octamers KVFGRCEL, VFGRCELA, ..., MKRHGLDN belong to the uncleaved category. The cleaved category, however, is not shift invariant because we believe the cleaving to occur at one specificity site and not in nearby sites.

A machine learning algorithm is one that can learn from experience (observed examples) with respect to some class of tasks and a performance measure. Machine learning methods are suitable for molecular biology data due to the learning algorithm's ability to construct classifiers/hypotheses that can explain complex relationships in the data. Recently, several works have approached the HIV-1 protease specificity problem by applying techniques from machine learning. In [2] the authors used a standard feed-forward multilayer perceptron (MLP) to solve this problem, achieving

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 413–420, 2005.

an error rate of 12%. In [3] the authors confirm the result of [4][2] using the same data and the same MLP architecture, showing that a decision tree was not able to predict the cleavage as well as MLP. In [5] the authors showed that HIV-1 protease cleavage is a linear problem and that the best classifier for this problem is linear-support vector machines (L-SVM).

Multiclassifier systems [6][7][8] are special cases of approaches that integrate several data-driven models for the same problem. A key goal is to obtain a better composite global model, with more accurate and reliable estimates. In addition, modular approaches often decompose a complex problem into sub-problems for which the solutions obtained are simpler to understand, as well as to implement, manage and update

In this work we show that the HIV-1 problem can be effectively solved using our hierarchical classifier (HC) architecture: the new approach yields an error rate of 7%, which is very lower that the best previous approaches (9.1% using L-SVM). Moreover, we show that if a step of feature selection and a step of feature transformation is performed, the radial basis function-support vector machines (R-SVM) yields an error rate of 8.4%.

The rest of the paper is organized as follows, in section 2 a brief description of the methods combined and tested in this work is given, in section 3 the results of the experiments are discussed, in section 4 a new hierarchical classifier is proposed and detailed, in section 5 the results of the new approach are reported and finally, in section 6, we draw some conclusions.

2 System and Methods

In this section a brief description of the feature extraction methodologies, feature transformations, classifiers and ensemble methods combined and tested in this work is given.

Feature Extraction (FE):

- Peptide sequences (PS):

A protein sequence is made from combination of variable length of 20 amino acids ∑ = {A,C,D….V,W,Y}. A peptide (small protein) is denoted by **P** = *P*4*P*3*P*2*P*1*P*1*'P*2*'P*3*'P*4*'* , where P_i is an amino-acid belonging to ∑. The scissile bond is located between positions *P*1 and *P*1'.

- Orthonormal encoding (OE):

It is the standard procedure [5] to map the sequence **P** to a sparse orthonormal representation. Each amino acid P_i is then represented by a 20 bit vector with 19 bits set to zero and one bit set to one, and each amino acid vector is orthogonal to all other amino acid vectors. P_i can take on any one of the twenty amino acid values.

Feature Selection (FS):

- Mahalanobis Ranking (MR):

We select the most discriminate features using a simple feature ranking which sorts the features that maximize the distance between the centroids of the different classes of patterns.

For each feature *k* we compute the scalar Mahalanobis distance ($dM_{i,k}$) [9] between the mean of the training patterns of each class *i* and the set of all training patterns. Features are then ranked according to the following separability measure:

$$\mathbf{DistF}(k) = \sum_{i=1}^{2} \sum_{j=1}^{2} \left| dM_{i,k} + dM_{j,k} \right|$$

In this paper we retain only the 100 features corresponding to the higher values of **DistF**.

Feature Transformation (FT):

- Karhunen-Loeve Transform (KL) [9]:

This transform projects high dimensional data onto a lower *k*-dimensional subspace (in this paper *k*=50) in a way that is optimal in a sum-squared sense. It is well known that KL is the best linear transform for dimensionality reduction.

Classifiers:

- SVM [9]:

The goal of this two-class classifier is to establish the equation of a hyperplane that divides the training set leaving all the points of the same class on the same side, while maximizing the distance between the two classes and the hyperplane.

- Edit distance classifier (EDC) [10]:

The edit distance of two strings, *s1* and *s2*, is defined as the minimum number of point mutations required to change *s1* into *s2*, where a point mutation is a change, insertion or deletion of a letter. The edit distance is coupled with a nearest-neighbor classifier in order to classify a new pattern.

Pattern Motifs [5]:

Studying the peptide sequence we can note that particular combinations of amino-acids influence the cleaving/non-cleaving decision. (i.e. if the third amino-acids of the peptide sequences is Glutamine we know that there is a low possibility that this peptide is a cleavage site). Pattern motifs can be used for creating a rule based classifier.

Multiclassifier Systems (MCS) [6]:

Multiclassifier systems are special cases where different approaches are combined to resolve the same problem. They combine, by a Decision Rule, output of various classifiers trained using different datasets.

3 Experimental Comparison

All the tests have been conducted on the following dataset:

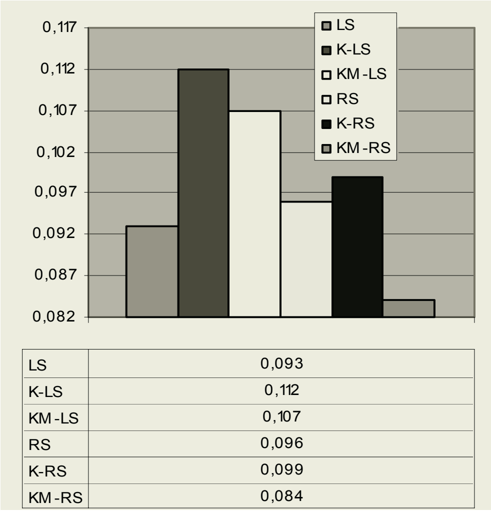
HIV data set [5] – The data set contains 362 octamer protein sequences each of which needs to be classified as an HIV protease cleavable site or uncleavable site. On this dataset, we performed 10 tests, each time randomly resampling learning, and test sets (containing respectively half of the patterns), but maintaining the distribution of the patterns in the two classes. The results reported refer to the average classification accuracy achieved throughout the 10 experiments.

Accuracy

We report some useful tests on the error rate aimed to compare the quality of various methods in the HIV-1 protease problem. Table 1 lists the tests whose error rates are reported in figure 1. The absence of the feature transformation step indicates that the classification task is performed starting from the original features.

Table 1. Tests made in HIV-1 protease problem.

Short Name	Feature Extraction	Feature Selection	Feature transformation	Classifier
LS	OE	-	-	L-SVM
K-LS	OE	-	YES	L-SVM
KM-LS	OE	YES	YES	L-SVM
RS	OE	-	-	R-SVM
K-RS	OE	-	YES	R-SVM
KM-RS	OE	YES	YES	R-SVM



LS	0,093
K-LS	0,112
KM-LS	0,107
RS	0,096
K-RS	0,099
KM-RS	0,084

Fig. 1. Error rate for different classifiers.

These results confirm as already stated in [5] that using the orthonormal encoding as feature extractor, the HIV-1 protease cleavage site specificity can be solved efficiently by linear models.

4 A New Hierarchical Structure

We develop a hierarchical structure, in which each step is constituted by a module able to classify only a fraction of the patterns: the rejected patterns are given as input to the following steps. The classifier is composed by four steps:

- Edit distance classifier (EDC) + Cleavage Rule (CR)
- 1°R-SVM
- Cleavage / Non Cleavage Rule (CNCR)
- 2°R-SVM

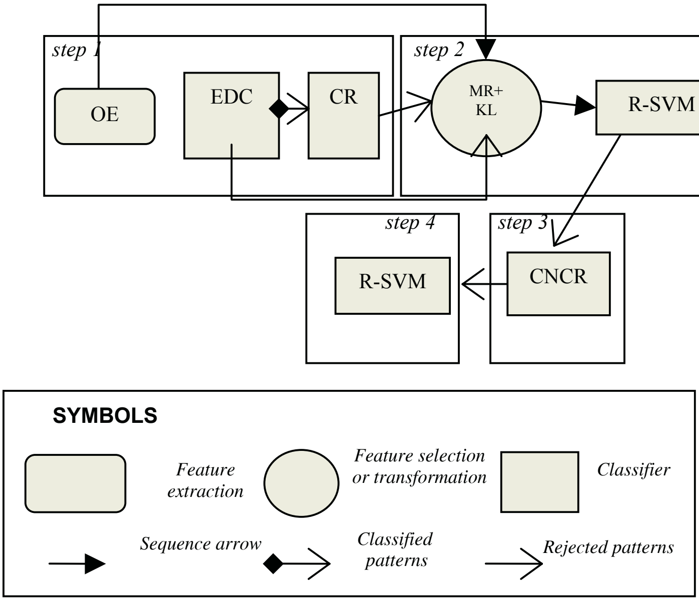


Fig. 2. Hierarchical classifier schema.

In figure 2 a detailed description of the system proposed is given.

Edit Distance Classifier + Cleavage Rule

The edit distance classifier gives good performance for a pattern belonging to the shift invariant class, while it is not reliable when assigns a pattern to the shift variant class. For example given a training set of patterns belonging to both the classes, if a new pattern is near (with respect the edit distance) to a pattern of the uncleaved class (shift invariant) we can reasonable assume that it belongs to the same class, on the contrary if the new pattern is near to a pattern of the cleavage class, we can not make any assumption with high degree of certainty.

Starting from this consideration we design a classifier that assign to the uncleaved class the patterns classified as uncleavage site by EDC, while reject the others.

The error rate of the EDC, if used without rejection to classify all the patterns, is approximately 84.20%. If we reject all the pattern assigned to the cleaved class, it is able to classify the 62.70% of patterns with an error rate of only 4.4%.

A possible method to reduce the error rate of edit distance classifier previous step is to reject the patterns, classified as uncleaved by the EDC, that satisfy thiese rules:

(xxx(NYLA)xxxx & (!xxxKxxxx | !xx(FKQ)xxxxx | !xxxxxCxx | !xxxxxxKx))

The rationale of this rule is:

- From a statistical study on the training set we have noted that a cleaved pattern with high similarity to an uncleaved one often contain the motif xxx(NYLA)xxxx (xxx(NYLA)xxxx means that the fourth amino-acid must be N,Y,L or A). This rule matches partially with a rule shown in [11].

- To avoid rejection of many patterns we use some motifs [5] that characterize the uncleaved class. These motifs are:
  - xxxKxxxx
  - xx(FKQ)xxxxx
  - xxxxxCxx
  - xxxxxxKx

By coupling these rules to the EDC we reject a further 20.6% of the patterns previously classified: this allows to reduce the error rate to 1.1% on the accepted patterns (which are the 49.83% of the total).

1° R-SVM

We adopt, as second step, the classifier denoted as KM-RS, in table 1: the patterns whose confidence is lower than a prefixed threshold are rejected. In this step the 31.6% of the patterns are classified, with an error rate of 7.3%.

Cleavage / Non-cleavage Rule

The third step consists in some rules proposed in [5]: if a pattern contains one of the motifs shown in table 2, it is classified as cleaved, if it contains one of the motifs shown in table 3 it is assigned to the uncleaved class, otherwise it is rejected to the next step.

Table 2. Pairs of amino acids and positions that influence the cleaving decision.

xxxFxExx	xxxYxExx	xxxLxExx	xxxFxQxx
xxVxxExx	xxxxPExx	xxVFxxxx	xxxFPxxx
xxIxxExx	xxxMxExx	xxAxxExx	xxAFxxxx
xxxxxExK	FxxxxExx		

Table 3. Single amino acids and positions that influence the cleaving decision.

xxxKxxxx	xxxxxSxx	xxxxxKxx	xxxPxxxx
xxxxCxxx	Cxxxxxxx	xxYxxxxx	

Using these rules, the 6.1% of the patterns can be classified, with an error rate of 3.6%.

2° R-SVM

The patterns rejected by the previous steps are finally classified by a R-SVM classifier. In this last step, all the remaining patterns are classified without rejection, with an error rate of 27%.

5 Results

In table 4 the classification performance of each step of the new hierarchical classifier are summarized: the local error rate is evaluated considering only the patterns effectively classified at each step, while the global error rate is the cumulative error obtained considering all patterns classified till that step; analogously with "local classified" we mean the percentage of the whole patterns classified at each step, while "global classified" is the cumulative percentage of classified patterns at each step. It is interesting to note that the 12.5% of patterns can be considered "difficult", since they contribute to generate the higher part of the total error rate.

Table 4. Error rate and number of patterns rejected at each step.

Steps	Global Error Rate	Local Error Rate	Global classified	Local classified
1	1.1	1.1	49.83	49.83
2	3.5	7.3	81.44	31.6
3	3.5	3.6	87.5	6.1
4	6.85	27	100	12.5

6 Conclusion

The problem addressed in this paper is to recognize, given a sequence of amino-acids, HIV-1 protease cleavage site. This is done by means of a hierarchical classifier that combines classifiers and rules based on pattern motifs. The major advantage of the proposed approach is the low error rate, better than other 'stand-alone' methods proposed in the literature. The major disadvantage is that system can not help to understand the relationship between the data.

As a future work we plan to improve the use of motifs rules and to develop a method similar to AdaBoost [13]. Moreover, some preliminary results [12] suggest the possibility of adopting editing methods to advance the classification performance in this field.

References

1. Beck, Z. Q.., Hervio, L., Dawson, P.E., Elder, J. E. and Madison, E.L.: Identification of efficiently cleaved substrates for HIV-1 protease using a phage display library and use in inhibitor development. Virology (2000).
2. Thompson, T.B., Chou, K.C. , Zheng, C.: Neural network prediction of the HIV-1 protease cleavage sites. Journal of Theoretical Biology, 177 (1995) 369-379.
3. Cai, Y.D., Chou, K.C.: Artificial neural network model for predicting HIV protease cleavage sites in protein. Advances in Engineering Software, 29 (1998) 119-128.
4. Narayanan, A., Wu, X., Yang, Z.: Mining viral protease data to extract cleavage knowledge. Bioinformatics, 18 (2002) 5-13.
5. Rögnvaldsson, T., You, L.: Why Neural Networks Should Not be Used for HIV-1 Protease Cleavage Site Prediction. Bioinformatics (2003) 1702-1709.
6. Dietterich, T.G.: Ensemble methods in machine learning. In J. Kittler and F. Roli, editors, Multiple Classifier Systems. First International Workshop, MCS 2000, Cagliari, Italy, volume 1857 of Lecture Notes in Computer Science, (2000) 1–15..
7. Masulli, F., Valentini, G.: Comparing decomposition methods for classification. In R.J. Howlett and L.C. Jain, editors, KES'2000, Fourth International Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies, (2000) 788–791.

8. Mayoraz, E., Moreira M.: On the decomposition of polychotomies into dichotomies. In The XIV International Conference on Machine Learning, (1997) 219–226.
9. Duda, R., Hart, P., Stork D.: Pattern Classification, Wiley, New York, 2001.
10. Levenshtein V. I.: Binary codes capable of correcting deletions, insertions and reversals. Doklady Akademii Nauk SSSR (1965) 845-848.
11. Tozser, J., Zahuczky, G., Bagossi, P., Louis, J., Copeland, T., Oroszlan, S., Harrison, R., Weber, T.: Comparison of the substrate specificity of the human T-cell leukemia virus and human immunodeficiency virus proteinases. European Journal of Biochemistry, (2000) 6287-6295.
12. Franco, A., Maltoni, D., Nanni, L.: Reward-Punishment Editing to appear on proceedings International Conference on Pattern Recognition (ICPR04), Cambridge (United Kingdom), 2004.
13. Houle, G., Aragon, D., Smith, R., Shridhar, Kimura, D.: A multilayered corroboration-based check reader. Document analysis system 2 (1998) 495-546.

Medial Grey-Level Based Representation for Proteins in Volume Images

Ida-Maria Sintorn[1], Magnus Gedda[2], Susana Mata[3,⋆], and Stina Svensson[1]

[1] Swedish University of Agricultural Sciences, Centre for Image Analysis,
Lägerhyddsvägen 3, SE-752 37, Uppsala, Sweden
ida,stina@cb.uu.se

[2] Uppsala University, Centre for Image Analysis,
Lägerhyddsvägen 3, SE-752 37, Uppsala Sweden
magnusg@cb.uu.se

[3] Rey Juan Carlos University, Dept. of Computer Science, Statistics and Telematics,
C/Tulipán s/n, 28933 Móstoles-Madrid, Spain
smata@escet.urjc.es

Abstract. We present an algorithm to extract a medial representation of proteins in volume images. The representation (MGR) takes into account the internal grey-level distribution of the protein and can be extracted without first segmenting the image into object and background. We show how MGR can facilitate the analysis of the structure of the proteins and thereby also classification. Results are shown on two types of protein images.

1 Introduction

Three-dimensional shape analysis is not an easy task. In many situations it is useful to analyze a simplified representation of the object, such as a skeleton instead of the original object. See for example [1–3]. A suitable representation can not only provide a more compact way to represent the object, which is of interest as volume images contain a large amount of data, but also bring out important features which are hardly noticeable in the original object. This is especially the case when the internal grey-level distribution of the object is of importance and is used for extracting a representation, [4–6].

To study the structure of proteins is central in molecular biology. The structure is often the key to understanding how flexible a protein is and how it can interact or bind to other proteins or substances, see, for example [7, 8]. However, imaging of a protein or a protein complex is a difficult task. Atomic resolution (angstrom scale) can be achieved through X-ray crystallography or nuclear magnetic resonance (NMR). Both have the drawbacks of being very time consuming and restricted to what type of proteins that can be imaged. Other faster and/or less restrictive methods are therefore of interest even if atomic resolution is not achieved. Using electron microscopy, two types of volume density images of protein structures can be acquired. Almost atomic level is possible to achieve by

[⋆] Performed the work while at Centre for Image Analysis, Uppsala, Sweden

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 421–428, 2005.

making a 3D reconstruction based on averaging from thousands of proteins of the same kind. Images in this resolution range have been used for structure analysis, e.g., in [6, 9, 10]. The individual particle imaging method called Sidec^{TM} Electron Tomography (SET), allows imaging at a resolution of approximately $2nm$, which is enough to reveal the main structural features of many proteins. This method also has the great benefits of allowing the study of proteins in solution or in tissue, as well as being fast in comparison with the other methods.

In this paper, a medial representation based on the internal grey-level distribution of an object is presented. The representation is a development of the method described in [4], adapted to the application of studying protein structure at the nm scale. The internal density distribution of a protein is of importance for its function and should, hence, be taken into account in a representation. Using the internal grey-level distribution also makes the representation independent of the initial segmentation provided that the object is connected. This is favorable as the objects at the resolution in consideration are very small and slight changes on the surface of the object, due to the choice of binarization method, can alter a representation based solely on outer shape significantly.

The representation scheme serves as a complement to the method described in [5]. There, a decomposition algorithm developed for the same type of images was presented, by which each protein is decomposed into simpler parts. In further analysis, it is of interest to follow the main structure connecting the different parts of the protein. Hence, a medial representation is of interest. The proposed representation can roughly be described as the maximum intensity path connecting center points of different parts of an object. The center points are identified as stable local maxima in the same manner as in the decomposition method presented in [5]. This compact representation facilitates comparison of different proteins as well as structural differences between proteins of the same kind. It gives information of how different subparts are connected and how the maximum internal density changes along a connection between two parts, revealing how tightly attached they are. As the representation is based on the internal density distribution, it gives information about cavities and tunnels unlikely to be detected by studying the shape of the binarized object.

2 Image Data

The representation scheme has been applied to two types of volume images: noise free density images constructed from a protein's atom positions deposited in the protein data bank (PDB) [11]; and SET images of proteins in solution. From a PDB entry, a volume image, where the grey-levels depict density, can be generated by placing a gauss kernel at each atom position and multiply by the mass of that atom [12]. The total density, or grey-level, in a voxel is then calculated by adding the contributions from gauss kernels of atoms in the vicinity of the voxel. This results in an image with floating point values, which is linearly stretched and rounded off to an 8-bit integer image. In our density reconstructions, we have used a σ of 1 leading to a resolution of $2nm$, which is the approximate

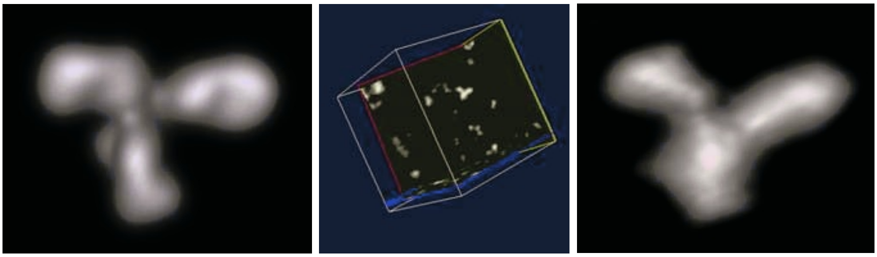


Fig. 1. Left: A volume rendering of the density reconstruction of an antibody, PDB ID 1igt. Middle: A cross section of a SET volume with an antibody almost in the middle. Right: A volume rendering of the antibody shown in the SET image.

resolution achieved with SET. To the left in Fig. 1, a volume rendered density reconstruction of an antibody (PDB ID 1igt) is shown.

The SET images are generated as follows. In the electron microscope, 2D projections of the flash frozen sample – in our case a solution containing the antibodies IgG [13] – are collected at several tilt-angles. These projections are then used to reconstruct the 3D sample by filtered back projection and a refinement method denoted Constrained Maximum Entropy Tomography, COMET [14]. The resulting SET volumes contain floating point data. They were converted to 8-bit integer data by linearly stretching all relevant values after a logarithmic transformation. A cross section of a SET image is shown in the middle of Fig. 1. The cutting plane was chosen to cut through an antibody, seen almost in the middle of the image. A volume rendering of the same antibody, extracted from the larger reconstruction volume, is shown to the right in Fig. 1. As can be seen in the middle of Fig. 1, a reconstruction volume contains lots of objects that are not true proteins. The SET proteins used for evaluating this method were cut from the reconstruction volumes after having been visually judged by an expert as being objects of interest. The SET images contain noise and they were therefore preprocessed by applying a Gauss filter with σ=2, which removes irrelevant local maxima.

3 Medial Representation Based on Internal Grey-Level Distribution

The proposed representation will in the following be denoted MGR for Medial Grey-level based Representation.

We consider grey-level images with grey-levels g_0 to g_n. A *region* R with grey-level g_k is a maximal connected set of voxels having grey-level g_k. A *cavity* is a region R, whose adjacent regions have grey-levels greater than g_k. The *background* consists of all regions with grey-level g_0 together with all cavities. The *object* is the union of all regions with grey-level g_k, $k = 1, \ldots, n$. We choose 26-connectedness for the object (protein) and, hence, 6-connectedness for the background. The object has a *tunnel* if there exists a background path passing through it.

To identify MGR, iterative thinning of the object is performed starting from voxels with grey-level g_1. The algorithm is a topology preserving algorithm as only *simple* voxels, [15], are considered for removal. This is important as MGR needs to be topologically equivalent to the object it represents in order to be a useful representation.

Within each grey-level g_k, distance information is used to guide thinning as it simplifies the identification of the current border of the object, in which voxels are considered for removal. For a region R with grey-level g_k the use of distance information will allow, if R is to be completely removed due to the existence of a neighboring region with higher intensity, thinning to continue towards higher grey-levels more internal in the object or, if R corresponds to a plateau, to center MGR in R.

For each iteration, removal is done in two scans: 1) border voxels that are not local maxima are marked for removal if they are simple; 2) marked voxels are sequentially removed if still simple. The result after this process is a medial structure consisting of a set of not necessarily one-voxel thick curves. This thinning process is repeated until no further changes occur resulting in MGR, a one-voxel thick curve structure.

As a post processing, we remove all cavities and tunnels which can be considered as spurious protrusions. A protruding cavity/tunnel is a cavity/tunnel which is connected to the main curve, i.e., the curve connecting local maxima, with only one branch. For the examples shown in this paper no such spurious protrusions occurred.

The resulting MGR reflects the topology of the initial object. Hence, if the object has a cavity, it will give rise to a cavity also in MGR. Moreover, if the initial object contains a tunnel, it will result in a loop in MGR. An illustration of this is shown for the 2D case in Fig. 2. There, a synthetic image resembling a slice of a protein (PDB ID 1d2t) is shown, left, and its grey-level landscape, right. The height in position (x_i, y_i) is equal to the grey-level for pixel (x_i, y_i). Three local maxima (red in the electronic version of the paper) were detected and connected through the ridges (blue) in the landscape. A local minimum is seen in the center (the pit in the landscape). This gives rise to a loop in MGR. Depending on the 3D structure of the protein, this will correspond to a tunnel or a cavity in the 3D image. For the protein PDB ID 1d2t, it corresponds to a cavity, as will be clear in the following Section.

4 Results

In Fig. 3 and the first row of Fig. 4, MGR for volumes constructed from PDB are shown. Some of the protein IDs are marked with a star and have an additional number. This is to point out that the protein is the assumed biologically functioning structure consisting of a number of entities of the protein. See for example the two last rows of Fig. 3 with the protein PDB ID 1eo8 and its as-

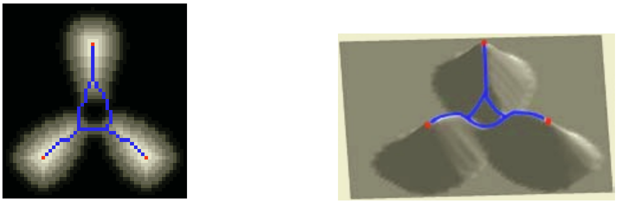


Fig. 2. Left: A synthetic 2D grey-level image. Right: Grey-level landscape. MGR and local maxima are overlayed (blue and red, respectively, in the electronic version of the paper).

sumed biological structure consisting of three 1eo8 units. In the last three rows of Fig. 4, MGR for IgG proteins imaged with SET is shown. Each row in Fig. 3 and 4 contains information regarding one protein object. In the first column, an identification (ID used in PDB or a numbering for the SET IgG proteins) is found. The proteins are visualized using volume rendering in column two. The third column shows their subparts [5], and the fourth their MGR. The number of local maxima, equaling the number of subparts, is found in column five, and finally, the number of endpoints in MGR in column six. Note that the number of endpoints is not necessarily equal to the number of local maxima. This is for example the case for ID 1afv and the assumed biological structure of ID 1d6i. The number of local maxima corresponds to the number of subparts in the object, while the number of endpoints give information about how these subparts are connected in MGR.

MGR can help to overcome many of the obstructions encountered when analyzing SET reconstructions of proteins. The first, and often most crucial step in image analysis, is to divide the image into objects and background. Due to noise and varying background, it is in most cases difficult to find good and reliable segmentation. Using MGR removes to a large extent the task of finding a reliable segmentation method. It is also difficult to decide what objects, out of the thousands present in a single reconstruction, are true proteins and not. MGR for a protein deposited in PDB can serve as a template for comparison of MGRs constructed from the objects present in the SET reconstructions. As long as an object is still connected, and the more intense inner parts are not removed, MGR is independent of small changes in the outer shape and thereby also the positioning of the border of the object.

From MGR, it is also possible to study how tightly different parts of an object are connected. If two parts are weakly connected, this can be seen as low intensities along MGR between them. If two parts are strongly attached, the intensities along MGR between them. An example is shown in Fig. 5. MGR of PDB ID 1afv, see also the third row of Fig. 3, is shown to the left, together with a plot of the grey-levels for the voxels along MGR to the right. It is clearly seen from the grey-levels along MGR that the the two main parts, each in turn consisting of two parts, are not very tightly connected, while the two parts making up each main part are very tightly connected.

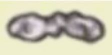
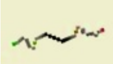
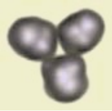
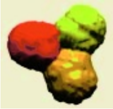
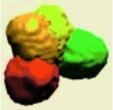
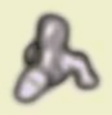
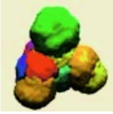
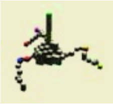
Id.	Vol. Rend.	Decomp.	MGR	Local max	Endpoints
1a2k				4	3
1afa				3	3
1afv				4	2
1axk				4	3
1ex4				3	3
1d2t* 3				3	3
1d6i				2	2
1d6i* 2				4	2
1eo8				4	4
1eo8* 3				12	4

Fig. 3. From left to right: identification; volume rendered original object; grey-level decomposition; MGR; number of local maxima; and number of endpoints in MGR for PDB volumes.

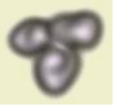
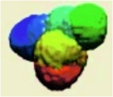
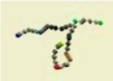
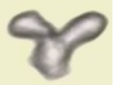
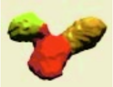
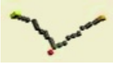
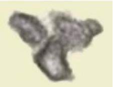
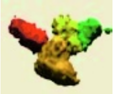
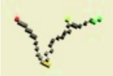
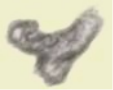
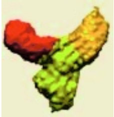
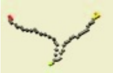
Id.	Vol. Rend.	Decomp.	MGR	Local max	Endpoints
1igt (PDB)				7	3
SET IgG No 1				3	3
SET IgG No 2				4	4
SET IgG No 3				3	3

Fig. 4. From left to right: identification; volume rendered original object; grey-level decomposition; MGR; number of local maxima; and number of endpoints in MGR for an antibody (1igt) from PDB and three IgG proteins from SET.

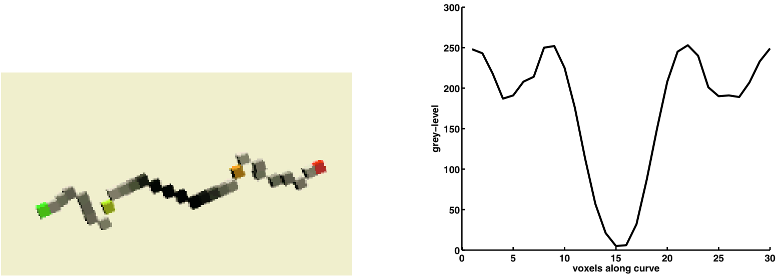


Fig. 5. Left: MGR of PDB ID 1afv. Right: Grey-levels along MGR.

MGR can also be of use when analyzing structural variations of a specific protein. We believe that information, for example on how flexible a protein in solution is, can be estimated by measuring angles between different parts of MGR from different SET images of the same protein. This would be much faster than fitting a model prior to making the measurements [13]. Developing stable methods to derive such information is left for future work.

Acknowledgement

We would like to thank Sara Sandin, the Karolinska Institute, for providing the SET volumes and the positions for the true proteins of interest. We would also like to thank the research staff at Sidec Technologies AB for valuable discussions.

References

1. Nyström, I., Smedby, Ö.: Skeletonization of volumetric vascular images – distance information utilized for visualization. J. Comb. Opt. **5** (2001) 27–41
2. Fouard, C., Malandain, G., Prohaska, S., Westerhoff, M., Cassot, F., Marc-Vergnes, J.P., Mazel, C., Asselot, D.: Skeletonization by blocks for large 3D datasets: Application to brain microcirculation. In Proc. 2nd. IEEE Int. Symp. Biomed. Imag. (2004) 89–92
3. Bitter, I., Kaufman, A.E., Sato, M.: Penalized-distance volumetric skeleton algorithm. IEEE Trans. on Vis. and Comp. Graph. **7** (2001) 195–206
4. Svensson, S., Nyström, I., Arcelli, C., Sanniti di Baja, G.: Using grey-level and distance information for medial surface representation of volume images. In Proc. 16th Int. Conf. on Pat. Rec. (ICPR 2002). Vol. 2., IEEE CS (2002) 324–327
5. Sintorn, I.M., Mata, S.: Using grey-level and shape information for decomposing proteins in 3D images. In Proc. 2nd. IEEE Int. Symp. Biomed. Imag. (2004) 800–803
6. De-Alarcón, P.A., Pascual-Montano, A., Gupta, A., Carazo, J.M.: Modeling shape and topology of low-resolution density maps of biological macromolecules. Biophys. J. **83** (2002) 619–632
7. Norel, R., Petrey, D., Wolfson, H.J., Nussinov, R.: Examintaiton of shape complementarity in docking of unbound proteins. Prot. Str. Func. Gen. **36** (1999) 307–317
8. Liang, J., Edelsbrunner, H., Woodward, C.: Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. Prot. Sci. **7** (1998) 1884–1897
9. Edelsbrunner, H.: Biological applications of computational topology. In Goodman, J.E., O'Rourke, J., eds.: Handbook of Discrete and Computational Geometry. CRS Press (2004)
10. Jiménez-Lozano, N., Chagoyen, M., Cuenca-Alba, J., Crazo, J.M.: Femme database: topologic and geometric information of macromolecules. J. Struct. Biol. **144** (2003) 104–113
11. Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalova, I., Bourne, P.: The protein data bank. Nucl. Ac. Res. **28** (2000) 235–242
12. Pittet, J.J., Henn, C., Engel, A., Heymann, J.B.: Visualizing 3D data obtained from microscopy on the internet. J. Struct. Biol. **125** (1999) 123–132
13. Sandin, S., Öfverstedt, L.G., Wikström, A.C., Wrange, O., Skoglund, U.: Structure and flexibility of individual immunoglobulin g molecules in solution. Structure **12** (2004) 409–415
14. Skoglund, U., Öfverstedt, L.G., Burnett, R., Bricogne, G.: Maximum-entropy three-dimensional reconstruction with deconvolution of the contrast transfer function: A test application with adenovirus. J. Struct. Biol. **117** (1996) 173–188
15. Bertrand, G., Malandain, G.: A new characterization of three-dimensional simple points. Pat. Rec. Let. **15** (1994) 169–175

Part VI

Medical Imaging

Automatic Classification of Breast Tissue

Arnau Oliver[1], Jordi Freixenet[1], Anna Bosch[1],
David Raba[1], and Reyer Zwiggelaar[2]

[1] Institute of Informatics and Applications, University of Girona, Spain
{aoliver,jordif,aboschr,draba}@eia.udg.es
[2] Department of Computer Science, University of Wales, Wales, UK
rrz@aber.ac.uk

Abstract. A recent trend in digital mammography are CAD systems, which are computerized tools designed to help radiologists. Most of these systems are used for the automatic detection of abnormalities. However, recent studies have shown that their sensitivity is significantly decreased as the density of the breast is increased. In addition, the suitability of abnormality segmentation approaches tends to depend on breast tissue density. In this paper we propose a new approach to the classification of mammographic images according to the breast parenchymal density. Our classification is based on gross segmentation and the underlying texture contained within the breast tissue. Robustness and classification performance are evaluated on a set of digitized mammograms, applying different classifiers and leave-one-out for training. Results demonstrate the feasibility of estimating breast density using computer vision techniques.

1 Introduction

Breast cancer is considered a major health problem in western countries, and indeed it constitutes the most common cancer among women. A study developed in 1998 by the American Cancer Society estimates that in western cultures between one in eight and one in twelve women will develop breast cancer during their lifetime. Breast cancer remains the leading cause of death for women in their 40s in the United States [1]. However, although breast cancer incidence has increased over the past decade, breast cancer mortality has declined among women of all ages. This favorable trend in mortality reduction may relate to the widespread adoption of mammography screening, in addition to improvements made in therapy [1].

Mammography remains the key screening tool for breast abnormalities detection, because it allows identification of tumour before being palpable. In a recent study, Vacek et al. [2] show that the proportion of breast tumours that were detected in Vermont by screening mammography increased from 2% during 1974 – 1984 to 36% during 1995 – 1999. However, of all lesions previously diagnosed as suspicious and sent for biopsy, approximately 25% were confirmed malignant lesions, and approximately 75% were diagnosed benign. This high false-positive rate is related with the difficulty of obtaining accurate diagnosis [3]. In this sense, computerized image analysis is going to play an important

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 431–438, 2005.

role in improving the issued diagnosis. Computer-Aided Diagnosis (CAD) systems are composed of a set of tools to help radiologists to detect and diagnose new cases. However, recent studies have shown that the sensitivity of these systems is significantly decreased as the density of the breast increased while the specificity of the systems remained relatively constant [4].

The origins of breast parenchymal classification are found in the work of Wolfe [5], who showed the relation between mammographic parenchymal patterns and the risk of developing breast cancer, classifying the parenchymal patterns in four categories. Since the discovery of this relationship automated parenchymal pattern classification has been investigated. Boyd et al. [6] proposed a semiautomatic computer measure based on interactive thresholding and the percentage of the segmented dense tissue over the segmented breast area. Karssemeijer [7] developed an automated method where features were computed from the grey-level values and the distance from the skin-line and are used in a k-Nearest Neighbour (kNN) classifier. Recently, Zhou et al. [8] proposed a rule-based scheme in order to classify the mammograms in four classes according to the characteristic features of the gray-level histograms.

A small number of previous papers have suggested texture representations of the breast. Miller and Astley [9] investigated texture-based discrimination between fatty and dense breast types applying granulometric techniques and Laws texture masks. Byng et al. [10] used measures based on fractal dimension. Bovis and Singh [11] estimated features from the construction of Spatial Gray Level Dependency matrices. Recently, Petroudi et al. [12] used textons to capture the mammographic appearance within the breast area. The approach developed by Blot and Zwiggelaar [13] is based on the statistical difference between local and median co-occurrence matrices computed over three regions of the breast. Related to this work, Zwiggelaar et al. [14] estimated the breast density by using co-occurrence matrices as features and segmentation based on the Expectation-Maximization algorithm. PCA was used to reduce the dimensionality of the feature space. This work was extended in [15] were a transportation algorithm was used for the feature selection process.

Our approach is also based on grouping those pixels with similar behaviour (gross segmentation), in our case consisting of similar tissue. Subsequently, texture features extracted from each region are used to classify the whole breast in one of the three categories that appears in the MIAS database [16]: fatty, glandular or dense breast. The remainder of this paper is structured as follows: Section 2 describes the proposed segmentation and classification method. Experimental results proving the validity of our proposal appear in Section 3. Finally, conclusions are given in Section 4.

2 Methodology

As we have mentioned in the introduction, previous work has used histogram information to classify breast tissue. However, in our experience and with our database, histogram information is not sufficient to classify the mammogram as

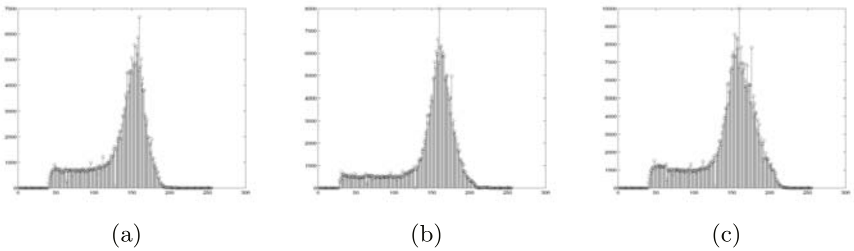


Fig. 1. Three similar histograms, each belonging to a different class of tissue. Concretely, (a) corresponds to a fatty breast, (b) to a glandular breast and (c) to a dense breast.

fatty, glandular or dense tissue. Figure 1 shows histograms for three different mammograms, each belonging to a different class.

Our approach is based on gross segmentation and the extraction of texture features of those pixels with similar tissue appearance of the breast. Using this set of features we train different classifiers and test them. But first of all, our approach begins with the segmentation of the profile of the breast.

2.1 Breast Profile Segmentation

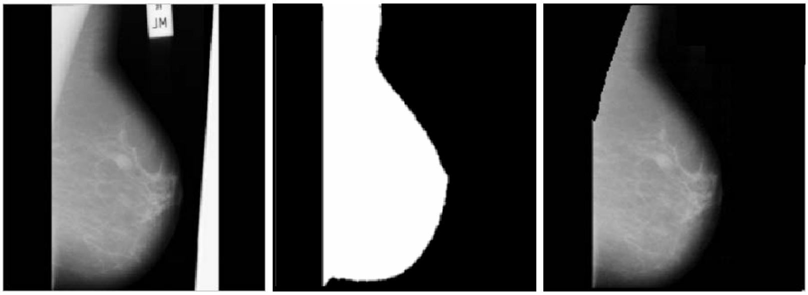
The segmentation of the foreground breast object from the background is a fundamental step in mammogram analysis. The aim of this process is to separate the breast from the rest of objects that could appear in a digital mammography: the black background; some annotations or labels; and the pectoral muscle.

In this work we used a previous developed algorithm [17] based on gray-level information. This algorithm begins by finding a threshold using histogram information, in order to separate the background from the rest of the objects of the mammogram, that is, the annotations and the union of the breast and pectoral muscle. The breast and pectoral muscle object is segmented looking for the largest object in the image. In order to separate the breast profile from the pectoral muscle we used an adaptive region growing approach, initializing the seed inside the pectoral muscle, and controlling the growing step using information about gray-level and the growth area.

Figure 2(a) shows a typical mammographic image. Applying the threshold and detecting the largest object, the union of the pectoral muscle and the breast area is found(Figure 2(b)). In the last image, the region of interest of the breast has been extracted from the pectoral muscle using the adaptive region growing algorithm described above.

2.2 Finding Regions with Similar Tissue

We consider that pixels from similar tissue have similar gray-level values, as can be seen in Figure 3. Hence, we use the k-Means algorithm [18] to group these pixels into separate categories. However, to avoid effects from microtexture that



(a) (b) (c)

Fig. 2. Sequence of the breast profile segmentation. (a) original image, (b) result of thresholding the image and detecting the largest region, and (c) segmented image without background and pectoral muscle.

could appear in some regions, we first smooth the breast region with a median filter.

The k-Means algorithm [18] is a popular clustering algorithm. It is defined as an error minimization algorithm where the function to minimize is the sum of errors squared:

$$e^2(K) = \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - c_k)^2 \tag{1}$$

where x_i are feature vectors, c_k is the centroid of cluster C_k, and K the number of clusters, which have to be known a priori. In our work, we selected $K = 2$ with the aim to obtain representative instances of two classes: fatty tissue and dense tissue.

When using the k-Means algorithm, the placement of the initial seed points plays a central role in obtaining the final segmentation results. Despite their importance, seeds are usually initialized randomly. In order to make a more informed decision, in our approach, the k-Means is initialized using histogram information. We initialized the two seeds with the gray level values that represent 15% and 85% of the accumulative histogram, with the objective to cluster fatty and dense tissue, respectively.

2.3 Extracted Features

Subsequent to k-Means, a set of features from the two classes that form the breast are extracted. In fact, a simple view of the feature space shows that using only the morphological features (like the centers of mass of both classes), is enough to distinguish between dense and fatty tissue. However, in order to distinguish glandular tissue, texture features needs to be considered. We used features derived from co-occurrence matrices [19].

Co-occurrence matrices are essentially two-dimensional histograms of the occurrence of pairs of grey-levels for a given displacement vector. Formally, the co-occurrence of grey levels can be specified as a matrix of relative frequencies

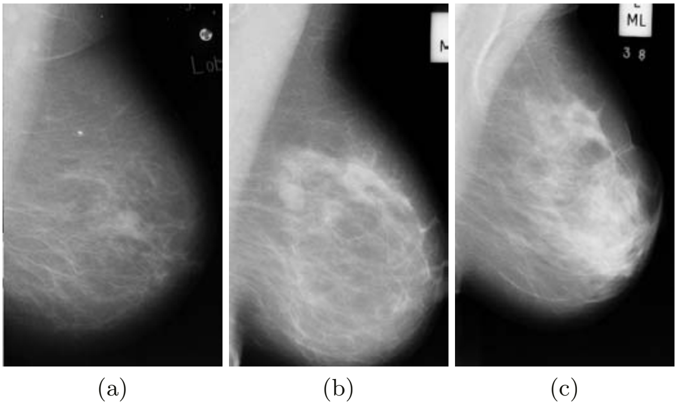


Fig. 3. Examples of different types of breast tissue in the MIAS database [16]. (a) fatty, (b) glandular, and (c) dense.

P_{ij}, in which two pixels separated by a distance d and angle θ have gray levels i and j. Co-occurrence matrices are not generally used as features, rather a large number of textural features derived from the matrix have been proposed [19]. Here we use 4 different directions: 0°, 45°, 90°, and 135°; and a distance equal to 1. For each co-occurrence matrix we determine the contrast, energy, entropy, correlation, sum average, sum entropy, difference average, difference entropy, and homogeneity features.

2.4 Classification

We evaluated two different kind of classifiers: the k-Nearest Neighbours algorithm and a Decision Tree classifier. The k-Nearest Neighbours classifier [20] (kNN) consists of classifying a non-classified vector into the k most similar vectors presents in the training set. Because kNN is based on distances between sample points in feature space, features need to be re-scaled to avoid that some features are weighted much more strongly than others. Hence, all features have been normalized to unit variance and zero mean.

On the other hand, a decision tree recursively subdivides regions in feature space into different subspaces, using different thresholds in each dimension to separates the classes "as much as possible". For a given subspace the process stops when it only contains patterns of one class. In our implementation we used the ID3 information criterion [20] to determine thresholds values from the training data.

3 Experimental Results

The method was applied on a set of 270 mammograms taken from the MIAS database, 90 of each class (fatty, glandular and dense). The spatial resolution of the images is 50μm x 50μm and the optical density is linear in the range

$0-3.2$ and quantized to 8 bits. To evaluate the method we performed three experiments.

The first experiment was performed over the set of fatty and dense mammograms, and using only morphological features extracted from the segmented clusters. We calculated the relative area, the center of masses and the medium intensity of both clusters. These features formed the input parameters for the classification stage. In order to evaluate the results, we used a leave-one-out method, in which each sample is analysed by a classifier which is trained using all other samples except for those from the same woman. The results showed that 87% and 82% of mammograms were correctly classified using the kNN classifier and the ID3 decision tree. However, when including the glandular class, both results were drastically decreased.

Table 1. Confusion matrices of (a) the k-NN classifier and (b) ID3 decision tree.

(a)

		Automatic Classification		
		Fatty	Glandular	Dense
Truth	Fatty	23	6	1
	Glandular	2	22	6
	Dense	1	14	15

(b)

		Automatic Classification		
		Fatty	Glandular	Dense
Truth	Fatty	18	10	2
	Glandular	4	21	5
	Dense	2	4	24

The second experiment was performed using 30 cases per class, and using the morphological features cited above as well the texture features. The efficiency of the classifiers were computed using again the same leave-one-out approach. Experimental results showed that classification results were improved when the cluster means were subtracted from the feature vectors. The reason for this can be found in the fact that increasing the dense area of the breast, results in a larger difference between the two tissue type clusters.

The confusion matrices for both classifiers are shown in Table 1. Confusion matrices should be read as follows: rows indicate the object to recognize (the true class) and columns indicate the label the classifiers associates at this object. As should be clear, the ID3 decision tree classifier has in general, a higher efficiency compared to the kNN classifier. This is due to the fact that the ID3 classifier contains a feature selection discrimination process. This ensures it avoids non-discriminant features to weight in the classification step. kNN classifiers do not have such feature selection and, therefore, the set of discriminant and non-discriminant features are weighted equally in the classification procedure. We can also note in Table 1 that mammograms belonging to fatty class are better classified than the rest of mammograms when using the kNN approach. On the other hand, dense mammograms are better classified by the ID3 approach.

The last experiment was performed by using a set of 90 mammograms per class, from which 30 were manually extracted from the set for training both classifiers, while the rest of mammograms constituted the testing set. The confusion matrices for this classification are shown in Table 2. Note that the ID3 classifier have again better results than the kNN. Moreover, it can be seen that the percentage of well-classified mammograms drastically decrease in comparative with the previous experiment. Concretely, in the leave-one-out method, the accuracy of the system was around 67% and 73% for the kNN and the ID3 classifiers respectively, while in the second experiment, the accuracy is about 56% and 61%. The main reason for this is likely to be the enlarged variance for the different mammographic tissue types.

Table 2. Confusion matrices of (a) the k-NN classifier and (b) ID3 decision tree.

(a)

		Automatic Classification		
		Fatty	Glandular	Dense
Truth	Fatty	38	19	3
	Glandular	9	33	18
	Dense	5	25	30

(b)

		Automatic Classification		
		Fatty	Glandular	Dense
Truth	Fatty	34	17	9
	Glandular	5	35	20
	Dense	2	18	40

4 Conclusions

This paper has presented an automatic classification method for the identification of breast tissue in mammographic images. The method is based on the integration of texture and gray level information. An initial method based on gray-level information starts segmenting the profile of the breast. Subsequently, the k-Means algorithm is used to segment the different tissue types of the mammograms. Morphological and texture features are extracted in order to characterize the breast tissue for each cluster. Finally, k-NN and ID3 are used to classify the breast as dense, fatty or glandular. Experimental results demonstrate the effectiveness of the proposed algorithm. Compared to published work, we can say that the developed method has a similar performances.

Further work will be focused on the characterization of the mammographic tissue in four classes, as are described in the Breast Imaging Reporting and Data System (BI-RADS). Moreover, other databases will be tested in order to evaluate in depth the efficiency of our proposal.

References

1. Buseman, S., Mouchawar, J., Calonge, N., Byers, T.: Mammography screening matters for young women with breast carcinoma. Cancer **97** (2003) 352–358
2. Vacek, P., Geller, B., Weaver, D., Foster, R.: Increased mammography use and its impact on earlier breast cancer detection in vermont. Cancer **94** (2002) 2160–2168
3. Basset, L., Gold, R.: Breast Cancer Detection: Mammograms and Other Methods in Breast Imaging. Grune & Stratton, New York (1987)
4. Ho, W., Lam, P.: Clinical performance of computer-assisted detection (cad) system in detecting carcinoma in breasts of different densities. Clinical Radiology **58** (2003) 133–136

5. Wolfe, J.: Risk for breast cancer development determined by mammographic parenchymal pattern. Cancer **37** (1976) 2486–2492
6. Boyd, N., Byng, J., Jong, R., Fishell, E., Little, L., Miller, A., Lockwood, G., Tritchler, D., Yaffe, M.: Quantitative classification of mammographic densities and breast cancer risk: results from the canadian national breast screening study. J. Natl Cancer Inst. **87** (1995) 670–675
7. Karssemeijer, N.: Automated classification of parenchymal patterns in mammograms. Physics in Medicine and Biology **43** (1998) 365–378
8. Zhou, C., Chan, H., Petrick, N., Helvie, M., Goodsitt, M., Sahiner, B., Hadjiiski, L.: Computerized image analysis: Estimation of breast density on mammograms. Medical Physics **28** (2001) 1056–1069
9. Miller, P., Astley, S.: Classification of breast tissue by texture and analysis. Image and Vision Computing **10** (1992) 227–282
10. Byng, J., Boyd, N., Fishell, E., Jong, R., Yaffe, M.: Automated analysis of mammographic densities. Physics in Medicine and Biology **41** (1996) 909–923
11. Bovis, K., Singh, S.: Classification of mammographic breast density using a combined classifier paradigm. In: International Workshop on Digital Mammography. (2002) 177–180
12. Petroudi, S., Kadir, T., Brady, M.: Automatic classification of mammographic parenchymal patterns: A statistical approach. In: International Conference of the IEEE Engineering in Medicine and Biology Society. Volume 2. (2003) 416–423
13. Blot, L., Zwiggelaar, R.: Background texture extraction for the classification of mammographic parenchymal patterns. In: Medical Image Understanding and Analysis. (2001) 145–148
14. Zwiggelaar, R., Blot, L., Raba, D., Denton, E.: Set-permutation-occurrence matrix based texture segmentation. In: Iberian Conference on Pattern Recognition and Image Analysis. (2003) 1099–1107
15. Zwiggelaar, R., Denton, E.: Optimal segmentation of mammographic images. In: International Workshop on Digital Mammography. (2004)
16. Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I., Boggis, C., Ricketts, I., Stamatakis, E., Cerneaz, N., Kok, S., Taylor, P., Betal, D., Savage, J.: The mammographic image analysis society digital mammogram database. In: International Workshop on Digital Mammography. (1994) 211–221
17. Raba, D., Oliver, A., Martí, J., Peracaula, M., Espunya, J.: Breast segmentation with pectoral muscle suppression on digital mammograms. In: Iberian Conference on Pattern Recognition and Image Analysis. (2005) to appear
18. MacQueen, J.: Some methods of classification and analysis of multivariate observations. In: Berkeley Symposium on Mathematical Statistics and Probability. Volume 1. (1967) 281–297
19. Haralick, R., Shanmugan, K., Dunstein, I.: Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics **3** (1973) 610–621
20. Duda, R., Hart, P., Stork, D.: Pattern Classification. 2 edn. John Wiley & Sons, New York (2001)

On Reproducibility of Ultrasound Image Classification*

Martin Švec[1], Radim Šára[1], and Daniel Smutek[2]

[1] Center for Machine Perception, Department of Cybernetics,
Faculty of Electrical Engineering, Czech Technical University Prague,
Technická 2, 166 27 Praha 6, Czech Republic
{xsvecm,sara}@cmp.felk.cvut.cz
http://cmp.felk.cvut.cz

[2] Charles University Prague, 1st Medical Faculty,
3rd Department of Medicine,
128 08 Praha 2, Czech Republic
smutek@cesnet.cz

Abstract. Ultrasound B-mode images of thyroid gland were previously analyzed to distinguish normal tissue from inflamed tissue due to Hashimoto's Lymphocytic Thyroiditis. This is a two-class recognition problem. Sensitivity and specificity of 100% was reported using Bayesian classifier with selected texture features. These results were obtained on 99 subjects at a fixed setting of one specific sonograph, for a given manual thyroid gland segmentation and sonographic scan orientation (longitudinal, transversal). To evaluate the reproducibility of the method, sensitivity analysis is the topic of this paper. A general method for determining feature sensitivity to variables influencing the scanning process is proposed. Jensen Shannon distances between modified and unmodified inter- and intra-class feature probability distributions capture the changes induced by the variables. Among selected features, the least sensitive one is found. The proposed sensitivity evaluation method can be used in other problems with complex and non-linear dependencies on variables that cannot be controlled.

1 Introduction

Hashimoto's lymphocytic thyroiditis (LT), one of the most frequent thyroid disorders, is a chronic inflammation of the thyroid gland. This disease changes the structure of the tissue. Changes are diffuse (they affect the entire gland) and can be detected by sonographic imaging. Information extracted from images by computers may provide additional support for diagnostic hypothesis.

Automatic recognition of LT has been attempted based on textural image features [12, 14]. Classification was done with features selected by a search pro-

* This work has been supported by the Grant Agency of the Czech Academy of Sciences under project 1ET101050403 and by the Czech Ministry of Health under project NO/7742-3.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 439–446, 2005.

cedure out of 129 features. The optimal features achieved sensitivity and specificity[1] of 100% in a cross-validation experiment on an independent set of 18 subjects [14].

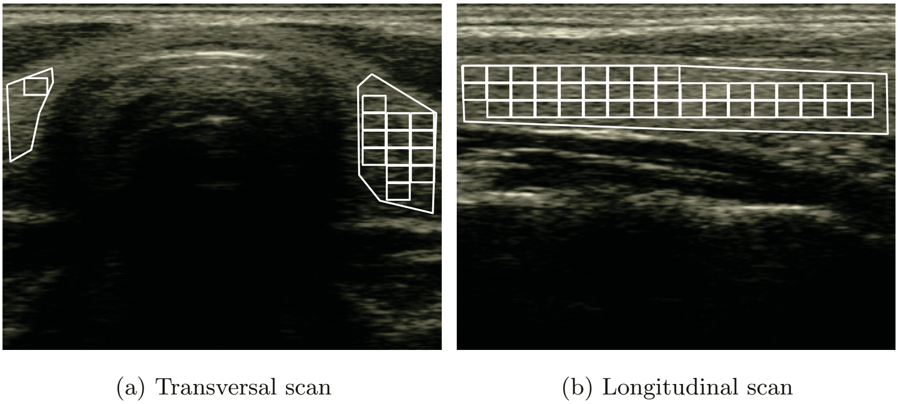
Although high success rate was achieved, the results were limited to one particular setting of one specific sonograph. This has been recognized as the most important obstacle to bringing the method to online clinical practice. The reproducibility issue is a long-standing problem in similar quantitative methods [4]. In relevant works, parameter settings were adjusted for optimal visualization [8], fixed to have standardized conditions [6, 11], or kept at values normally used in clinical practice [3]. Chan [1] tried to tackle this problem by changing the gain setting during the experiment and capturing for each gain at least five images of the object. Mojsilovic [9] removed the mean of each image in order to eliminate effects of unequal ultrasound gain settings.

The goal of this paper is to *quantify reproducibility* of features used previously [14]. Reproducibility is the possibility to achieve the same classification results under different sonograph setting, different gland delineations in the manual segmentation step (depending on physician's knowledge and experience), and different scan orientation (longitudinal or transversal). The proposed analysis is general enough to be applied to other data interpretation problems involving complex and non-linear dependencies on variables that cannot be controlled. The rest of this paper is structured as follows. Texture features are described and sensitivity analysis method is proposed in Sec. 2. Sec. 3 describes a feature sensitivity experiment and results. Discussion follows in Sec. 4 and conclusions are given in Sec. 5.

2 Methods

Given a sonographic B-mode image, a two-class classification problem is considered in the previous work [12, 14]: distinguishing healthy tissue (denoted here as N) from tissue changed due to Hashimoto's Lymphocytic Thyroiditis (denoted as LT). The classification is done on textural features computed from a set of fixed-size rectangular regions referred to as texture samples, as shown in Fig. 1. The non-overlapping samples are obtained from a manually segmented thyroid gland. In previous work, optimally performing features from the set of 129 candidates consisted of Haralick's texture features [5] and Muzzolini's spatial features [10] were automatically searched for. Their performance was measured as Bayes classifier error. The classifier was learned on a training set of 81 patients and classification error was evaluated on an independent test set of 18 subjects. Three one-dimensional features turned up, each for a different texture sample size, i.e. F2 for 41×41, F6 for 31×31 and F7 for 21×21 texture samples (features are named consistently with previous work [14]; the size is given in pixels). The selected features achieved sensitivity and specificity of 100% in a cross-validation experiment on the independent set of 18 subjects. The principal parameters of

[1] Sensitivity is the proportion of subjects with disease who have a positive test result, specificity is the proportion of subjects without disease who have negative test result.


(a) Transversal scan

(b) Longitudinal scan

Fig. 1. Sonographic images with manually segmented thyroid gland and covered by rectangular texture samples.

the sonograph[2] were fixed in the study: the gain of 92, medium sensitivity by depth, maximal acoustic power, frequency of 8MHz, repetition rate of 19Hz, and maximum spatial resolution of 4cm. All details concerning data acquisition and processing are given in [14].

Features used in the current sensitivity analysis are the selected features F2, F6 and F7, as described above. Feature probability distributions (FPD) for each class were estimated by histogramming. Optimal (the least bias and variance) histogram resolution according to Scott's rule was used [13]. The idea of the sensitivity analysis is to quantify the changes of these histograms under various modifications of the data acquisition.

A suitable statistic for this purpose is a divergence measure between two feature probability distributions. Jensen-Shannon divergence (*JS*) is used as a semi-definite, additive and symmetric measure. Let X be the range of a discrete random variable and let p_1 and p_2 be two probability distributions over X. The *JS* is defined in terms of discrete Shannon entropy $H(p)$ of a probability distribution function p as

$$JS(p_1, p_2) = H\left(\frac{p_1 + p_2}{2}\right) - \frac{H(p_1) + H(p_2)}{2} \, . \tag{1}$$

A detailed description is given in [7] and recommendations for practical usage in [2].

Using *JS* the sensitivity is measured by comparing the inter-class difference d_I (difference between N and LT class) and the within-class difference $d_{W|N}$ or $d_{W|LT}$ (difference between FPD and changed FPD, for given class N or LT). Changes in FPD are given by different 1) sonograph gain setting; 2) thyroid gland segmentation; and 3) scan orientation according to the following diagram

[2] Toshiba ECCO-CEE, console model SSA-340A, transducer model PLF-805ST.

$$\begin{tikzcd}
FPD|N \arrow[r, leftrightarrow, "d_{W|N}"] \arrow[d, leftrightarrow, "d_I"'] & FPD'|N \\
FPD|LT \arrow[r, leftrightarrow, "d_{W|LT}"] & FPD'|LT
\end{tikzcd} \tag{2}$$

where $FPD|N$ stands for 'FPD given class N' and $FPD'|N$ stands for the same under changed conditions. The inter-class distance d_I was measured on the full dataset (99 subjects) under the standard gain and the standard segmentation. If d_W is higher than d_I, the feature is sensitive on given changes and reproducibility can not be achieved.

3 Experiments and Results

Sensitivity of the features to sonograph gain setting was determined by computing the distance $d_{W|N}$ between the N class FPDs under the standard gain of 92 and the respective distributions obtained from a set of images from one subject (class N) under two other gain settings (90, 94). The result was compared to the inter-class distance d_I.

A similar method was used to assess the sensitivity of features to the thyroid gland segmentation. Three different boundary delineations for one subject of class N were drawn by another physician in addition to the one used in feature selection.

Finally, the influence of the scan orientation on the features was assessed. Given the feature, the distance between the longitudinal and transversal scans was measured in each class denoted here as $d^s_{W|N}$, $d^s_{W|LT}$, respectively. The larger of the two values $\max(d^s_{W|N}, d^s_{W|LT})$ was compared to the smaller of the two inter-class distances $\min(d_{I|\mathrm{long}}, d_{I|\mathrm{trans}})$ computed for each scan orientation separately according to the following diagram

$$\begin{tikzcd}
FPD|N, \mathrm{long} \arrow[r, leftrightarrow, "d^s_{W|N}"] \arrow[d, leftrightarrow, "d_{I|\mathrm{long}}"'] & FPD'|N, \mathrm{trans} \arrow[d, leftrightarrow, "d_{I|\mathrm{trans}}"] \\
FPD|LT, \mathrm{long} \arrow[r, leftrightarrow, "d^s_{W|LT}"] & FPD'|LT, \mathrm{trans}
\end{tikzcd} \tag{3}$$

The results of the sensitivity analysis under gain change are shown in Tab. 1. In 21 × 21 texture samples the differences due to varying gain setting are consistently smaller than the inter-class difference d_I. In 31 × 31 and 41 × 41 samples the differences (shown in bold) are already comparable to d_I.

Tab. 2 shows that changes due to the different segmentations s_1, s_2, s_3 are bigger than the inter-class difference d_I for all three features. Again, the least influenced are the 21 × 21 texture samples.

In Tab. 3 results for different scan orientation (longitudinal, transversal) are shown. The inter-class distances (last two rows) for 31 × 31 and 41 × 41 samples are consistently greater than the inter-scan distances (first two rows). Only in the 21 × 21 samples the two values (in bold) are comparable.

Table 1. The *JS* distances between selected features under the gains of 90 and 94 and the same features under the standard gain of 92 according to Eq. (2). The last row shows inter-class distances.

	F7, 21×21	F6, 31×31	F2, 41×41
$d_{W\|N,\ \mathrm{gain}=90}$	0.031	**0.573**	**0.542**
$d_{W\|N,\ \mathrm{gain}=94}$	0.158	0.409	0.475
d_I	0.225	0.532	0.510

Table 2. The *JS* distances between selected features computed under different thyroid gland segmentations s_1, s_2, s_3 and the same features under the standard segmentation according to Eq. (2).

	F7, 21×21	F6, 31×31	F2, 41×41
$d_{W\|N,s_1}$	0.208	**0.701**	**0.850**
$d_{W\|N,s_2}$	0.151	**0.859**	**0.888**
$d_{W\|N,s_3}$	**0.390**	**0.904**	**0.844**
d_I	0.225	0.532	0.510

Table 3. The *JS* distances between selected features from individual scan orientations (first and second row) as compared to the inter-class distances in individual scan orientations (third and fourth row) according to Eq. (3).

	F7, 21×21	F6, 31×31	F2, 41×41
$d^s_{W\|N}$	0.033	0.042	0.021
$d^s_{W\|LT}$	**0.252**	0.092	0.014
$d_{I\|\mathrm{trans}}$	0.478	0.276	0.409
$d_{I\|\mathrm{long}}$	**0.184**	0.575	0.389

An example of feature probability distributions for N and LT tissue used in this analysis are shown in Fig. 2. We can see the *JS* distances capture the relative differences between the histograms well.

Since gain setting is the parameter that has the greatest influence on visual appearance of sonographic image, features were extracted for a wider gain range (82–100). Measurements were done on a grey-scale phantom[3]. The scatter plots of feature $\mathrm{F7}_{92}$ under gain 92 and $\mathrm{F7}_g$ under several other gains g are shown in Fig. 3. One data point corresponds to one texture sample (21×21). It can be seen that points make clusters and the mapping induced by the gain change is too complex to be mathematically described. Next we considered simple feature F0 defined as the mean value over the whole rectangular texture sample. Analogical plots to those in Fig. 3 using F0 are shown in Fig. 4.

[3] Precision Small Parts Grey Scale Phantom Gammex 404GS LE.

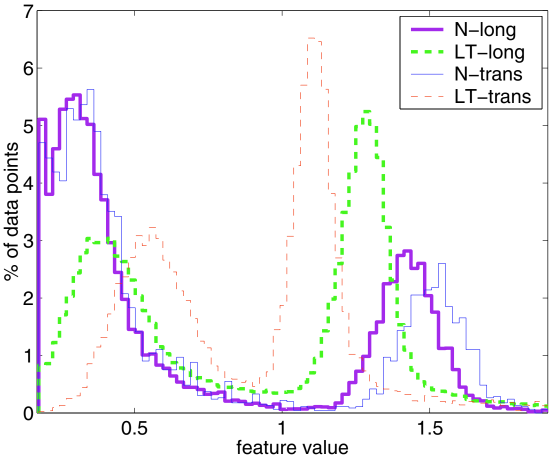


Fig. 2. Histograms for N and LT tissue (feature F7, 21 × 21 texture samples). We can see histograms for longitudinal scans are more similar than for transversal scans ($d_{I|\mathrm{long}}$ is smaller than $d_{I|\mathrm{trans}}$, see Tab. 3).

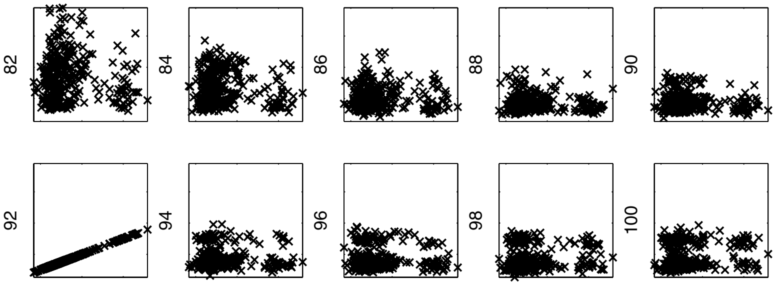


Fig. 3. Comparison of feature F7 under standard gain with F7 under gains of 82 to 100. Coordinates of each point are feature values for standard gain (92, x-axis) and under gain change (82-100, y-axis). 21 × 21 texture samples are used.

4 Discussion

Note that the image area of displayed thyroid tissue is much smaller in transversal scans than in longitudinal ones (see Fig. 1). This means that a smaller number of texture samples fit within the boundary of thyroid gland in a transversal scan as compared to a longitudinal scan. Larger texture samples do not cover the area of the transversal scans well. Therefore, substantial part of available information can be lost for feature construction process from transversal scans. Longitudinal scans provide greater amount of image data from a larger contiguous area of the gland tissue, therefore they should be more useful for automatic texture analysis. However, as can be seen from the last two rows of Tab. 3, distance between N and LT tissue is not always bigger for longitudinal scans than for transversal

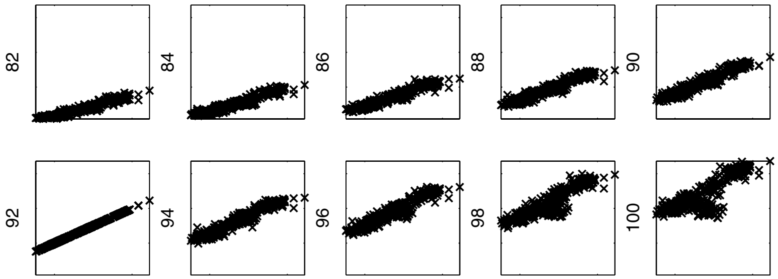


Fig. 4. Comparison of feature F0 under standard gain with F0 under gains of 82 to 100. Coordinates of each point are feature values for standard gain (92, x-axis) and under gain change (82-100, y-axis). 21 × 21 texture samples are used.

scans. This can be due to longitudinal artifacts in surrounding and examined tissue, e.g. muscle fibres or vessels. On the other hand we saw in Tab. 3 that inter-class distance is large in transversal scans when F7, 21 × 21 features are used and in longitudinal scans when F6, 31 × 31 features are used. Hence, the results could be improved by taking into account longitudinal and transversal images individually, e.g. by combining two classifiers, one using F7 on transversal scans and another using F6 on longitudinal scans.

There is high sensitivity to thyroid gland segmentation according to *JS* distance in larger samples (see Tab. 2). This can be related to sample placement method that leaves small areas along the boundaries uncovered by samples.

To guarantee reproducibility of results under different gain settings, transformation to recalculate features from arbitrary gain to standard gain should be found. From the results shown in Figs. 3,4 it follows that direct transformation of complex features is unfeasible but re-mapping of the raw image values prior to feature computation seems feasible. The results on F0 (see Fig. 4) reveal two components: one approximately linear and another random (see the cluster just below the linear cluster). The origin of this cluster is not known. Further analysis is necessary.

An initial calibration (e.g., using a gray-scale phantom) and subsequent customization of the recognition tool could be another approach for solving the problem of reproducibility. The phantom would need to be specially designed to reproduce the statistical distribution of those features that were found to be optimal for the LT/N classification task. Whether this is feasible remains to be ascertained.

5 Conclusions

The sensitivity analysis shows that the results for 31 × 31 texture samples and 41 × 41 texture samples are sensitive to small changes in sonograph setting. Both are also sensitive to different gland segmentations. They are stable under transversal and longitudinal scans.

The 21×21 pixel samples are insensitive to different gain settings and their sensitivity to different gland segmentations is small. They can also distinguish scan orientation, since there is a significant difference between inter-class distances of longitudinal and transversal scans. Distance between N and LT tissue is bigger for transversal than for longitudinal scans.

For greater difference in sonograph parameter setting it will be necessary to remap raw image values by a corrective transformation. We believe that if features of small sensitivity are used subsequently, the classification results will be reproducible. The corrective transformation is a topic for ongoing work.

References

1. Chan, K.L.: Adaptation of ultrasound image texture characterization parameters. In Proc Int Conf IEEE Eng in Medicine and Biology, vol. 2, pp. 804–807, 1998
2. Dhillon, I., Manella, S., Kumar, R.: Information theoretic feature clustering for text classification. Tech. Rep. TR–02–17, Dept of CS, U of Texas at Austin, USA, 2002
3. Dixon, K.J., Vince, D.G., Cothren, R.M., Cornhill, J.F.: Characterization of coronary plaque in intravascular ultrasound using histological correlation. In Proc Int Conf IEEE Eng in Medicine and Biology, volume 2, pages 530–533, 1997
4. Garra, B.S., Krasner, B.H., Horii, S.C., Ascher, S., Mun, S.K., Zeman R.K.: Improving the distinction between benign and malignant breast-lesions: The value of sonographic texture analysis. Ultrasonic Imaging, 15(4):267–285, 1993
5. Haralick, R.M.: Statistical and structural approaches to texture. In Proc IEEE, volume 67, pages 786–804, 1979
6. Hirning, T., Zuna, I., Schlaps, D., Lorenz, D., Meybier, H., Tschahargane, C., van Kaick, G.: Quantification and classification of echographics findings in the thyroid gland by computerized B-mode texture analysis. Europ J Radiol, 9(4):244–247, 1989
7. Lin, J.: Divergence measures based on the Shannon entropy. IEEE Trans Inform Theory, 37(1):145–151, 1991
8. Mailloux, G., Bertrand, M., Stampfler, R., Ethier, S.: Computer analysis of echographic textures in Hashimoto disease of the thyroid. J Clinical Ultrasound, 14(7):521–527, 1986
9. Mojsilovic, A., Popovic, M., Sevic, D.: Classification of the ultrasound liver images with the $2N$ multiplied by 1-D wavelet transform. In Proc IEEE Int Conf Image Processing , volume 1, pages 367–370, 1996
10. Muzzolini, R., Yang, Y., Pierson, R.: Texture characterization using robust statistics. Pattern Recognition, 27(1):119–134, 1994
11. Pohle, R., von Rohden, L., Fisher, D.: Skeletal muscle sonography with texture analysis. In Proc Medical Imaging, vol. 3034 of Proc SPIE, pp. 772–778, 1997
12. Šára, R. Švec, M., Smutek, D., Sucharda, P., Svačina, Š.: Texture analysis of sonographic images for diffusion processes classification in thyroid gland parenchyma. In Proc Conf Analysis of Biomedical Signals and Images, pages 210–212, 2000
13. Scott, D.W.: Multivariate Density Estimation. John Wiley, 1992
14. Smutek, D., Šára, R., Sucharda, P., Tardi, T., Švec, M.: Image texture analysis of sonograms in chronic inflammations of thyroid gland. Ultrasound in Medicine and Biology, 29(11):1531–1543, 2003

Prior Based Cardiac Valve Segmentation in Echocardiographic Sequences: Geodesic Active Contour Guided by Region and Shape Prior

Yanfeng Shang[1], Xin Yang[1], Ming Zhu[2], Biao Jin[2], and Ming Liu[2]

[1]Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University
Shanghai 200030, P.R. China
{aysyf,yangxin}@sjtu.edu.cn
[2]Xinhua Hospital, Attached to Shanghai Second Medical University
Shanghai 200092, P.R. China
zhuming58@vip.sina.com, {king1669,lesserniuniu}@hotmail.com

Abstract. This paper presents a segmentation of cardiac valve structure method in ultrasound sequence. Prior knowledge on certain complex object is a powerful guidance in image segmentation. We represent region and shape prior of the cardiac valve in a form of speed field and incorporate it into image segmentation process within level set framework. Region prior constrains the zero level set evolving in certain region and shape prior pulls the curve to the ideal contour. Experiments on a large quantity of 3D valve sequences show that the algorithm improves accuracy of segmentation and reduces the manual intervention.

1 Introduction

The algorithm presented in this paper was originally developed for a project of reconstruction of mitral valve leaflet from 3D ultrasound images sequences, which leads to a better understanding of cardiac valve mechanics and the mechanisms of valve failure. Furthermore segmenting valve efficiently could also aid surgery in diagnosis and analysis of cardiac valve disease.

Among medical imaging techniques, ultrasound is particularly attractive because of its good temporal resolution, noninvasiveness and relatively low cost. In clinical practice, segmentation of ultrasound images still relies on manual or semi-automatic outlines produced by expert physicians, especially when it comes to the object as complex as cardiac valve. And the large quantities of data of 3D volume sequences make a manual procedure almost impossible. Efficient processing of the information contained in ultrasound image calls for automatic object segmentation techniques.

As for the echocardiographic sequences, the inherent noise, similarity of intensity distribution between object and background, and complex movements, make it difficult to segment the cardiac valve accurately. We could make full use of the valve priors, and segment it under the guidance of prior knowledge to reduce the manual intervention. We can segment the valve automatically guided by the following prior:
1) The valve moves in a relatively fixed region between neighbor images.
2) The valve has a relatively predetermined shape at a certain position.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 447–454, 2005.

We developed an algorithm based on a level set framework, and represented the priors as speed fields which drive the zero level converging on the ideal contour. Section 2 of this paper gives an overview of some of the existing prior based object segmentation methods. The proposed algorithm is formulated in Section 3. Result and application are presented in Section 4 and conclusion follows in Section 5.

2 Review

Prior knoledge on an object can be a significant help in segmenting or locating process. Prior based image segmentation, which incorporates the prior information, allows the ambiguous, noise-stained or occluded contour clear and the final result becomes more robust, accurate and efficient. Snake and level set are the two models, which are used to incorporate prior knowledge to segment certain object.

There were some applications of prior based image segmentation under snake framework in the past several years. The snake methodology defines an energy function over a curve as the sum of an internal and external energy of the curve, and evolves the curve to minimize the energy [1]. Priors can be incorporated into the function as an energy item freely. D. Cremers [2,3] established a prior Point distribute Model (PDM) of certain object, then calculated the contour's post Bayesian probability and create an energy item to control the evolving process. I. Mikic[4] modified the internal energy to preserve the thickness of the cardiac valve leaflet. But under snake framework, it hardly to deal with topology changes since the object is described by a point serial. Unfortunately, there is great topology change during the valve opening and closing. So the snake model is too hard to employ in our project. And what is more, snake model needs an initial outline close to the contour, which makes an automatic segmentation process impossible.

Level set based segmentation embeds an initial curve as the zero level set of a higher dimensional surface, and evolves the surface so that the zero level set converges on the boundary of the object to be segmented. Since it evolves in a higher dimension, it can deal with topology changing naturally. The difficulty is how to incorporate prior knowledge into the evolution process. Leventon[5] calculated the similarity between the zero level and the statistical prior contours, and made it be an item of evolution equation. Chen[6] made the similarity metric applicable to linear transform. They all got a more robust result than traditional method. But calculating the similarity between zero level and prior statistical contour at each evolution step is time-consuming, and a scalar of contour similar metric can't guide the evolution directly. In the following, we present a new algorithm that represents the region and shape prior knowledge in a form of speed field. The speed vector could drive the zero level set directly to the ideal cardiac valve contour.

3 Geodesic Active Contour Guided by Prior Knowledge

Our prior guided valve segmentation is based on Geodesic Active Contour (GAC) [7, 8],

$$\frac{\partial\phi}{\partial t}=u(x)(k+v_0)|\nabla\phi|+\nabla u\cdot\nabla\phi \qquad (1)$$

$$Where: \qquad u(x)=-|\nabla G_\sigma * I|$$

where v_0 is an image-dependent balloon force that drives the contour to flow outward. In this level set framework, the surface φ evolves at every point perpendicular to the level sets as a function of the curvature at that point and the image gradient.

Prior knowledge could be a powerful guidance to zero level set evolution. We added new speed items to evolution equation of GAC. The new prior knowledge items form a total force with internal and external force of the original image and drive the zero level set converge to the ideal contour. There are two levels of additional prior force, one is the lower level of region prior constrain which make the zero level evolve in certain region, the other is the higher level of shape prior constrain which makes the final contour converges to the prior shape of object. Then we obtain a new equation:

$$\frac{\partial\phi}{\partial t}=u(x)(k+v_0)|\nabla\phi|+\nabla u\cdot\nabla\phi+\sum F_i\cdot\nabla\phi \qquad (2)$$

where F_i is the region, shape or the other prior knowledge forces. How to transform the prior knowledge into a speed field is presented in details in following.

3.1 Region Prior Based GAC

The movement of cardiac valve is very complex. It goes with beating of heart, turns around the valve root, and makes great deformation by itself. But the whole valve moves in a relatively fixed region, which is just in the ventricle. When it comes to 3D echocardiographic sequence, the valve could share the same region in different sample position of the same time. So does it in same position of different time. Then the whole 3D echocardiographic sequence could be segmented based on several prior regions that is just the ventricle or a manual outline. When the zero level set is limited to evolve in the fix region, the segmenting process will be more robust and efficient.

Consider a prior region Ω where the valve moves, define a region function J(x,y):

$$J(x,y)=\begin{cases}1 & (x,y)\in\Omega\\ 0 & (x,y)\notin\Omega\end{cases} \qquad (3)$$

A speed field is created outside the prior region. The force of the field is zero inside the prior region and direct to the prior region outside it. The power of the force has close relation to the distance from the point to the prior region. So the speed field has a potential to drive zero level set to the prior region. And the prior force would get a balance with the inflating force nearby the boundary of the region. When segmenting the valve, appropriate contour could be got at the root of valve lest the zero level set evolve to the whole cardiac wall. The distance from point X to prior region Ω is defined as:

$$\gamma(X)=\begin{cases} d(X) & X\notin\Omega \\ 0 & X\in\Omega \end{cases} \tag{4}$$

$$d(X)=\min(|X-X_I|) \qquad X_I\in\Omega \tag{5}$$

Then the speed field of the prior region is:

$$F_{region}(X)=[f_r(d)+c_1]\frac{\nabla\gamma}{|\nabla\gamma|} \tag{6}$$

where c_1 equal or a little less than v_0; $f_r()$ makes the prior force almost c_1 nearby Ω and rise to c_1+c_2 far away from Ω. We take:

$$f_r(d)=c_2\left(1-\exp\left(-\frac{d^2}{\sigma^2}\right)\right) \tag{7}$$

Then a speed field is gotten which could drive the zero level set to Ω.

The final region prior based GAC is:

$$\frac{\partial\phi}{\partial t}=u(x)(k+v_0)|\nabla\phi|+\nabla u\cdot\nabla\phi+[f_r(d)+c_1]\frac{\nabla\gamma}{|\nabla\gamma|}\cdot\nabla\phi \tag{8}$$

A final result could be gotten by some post-processing which includes the cardiac valve, the root of valve and the raised cardiac wall, which joggles with the valve. Erode the prior region Ω and get Ω', fill the segmentation result with foreground color outside Ω', make an intensity reversion, then the valve is gotten.

Pre-processing is very important for noises are inherent in the ultrasound image. Modified Curvature Diffusion Equation (MCDE)[9,10] is a excellent choice, which could preserve the edge and reduce the noise. The equation is given as:

$$u_t=|\nabla u|\nabla\cdot c(|\nabla u|)\frac{\nabla u}{|\nabla u|} \tag{9}$$

An example of cardiac valve segmented by region prior based GAC is given in Fig. 1.

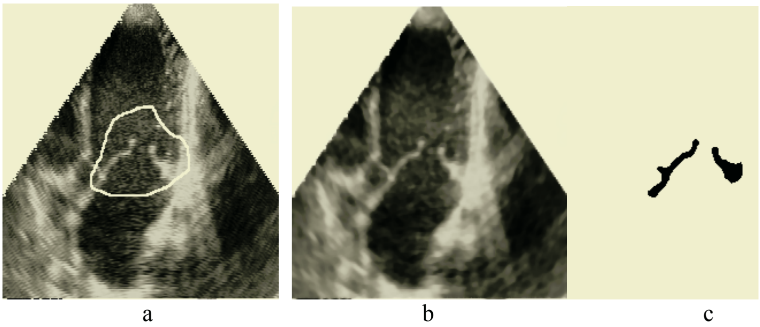


Fig. 1. Cardiac valve segmented by region prior based GAC. (a) An image from echocardiographic sequence and prior valve region. (b) Result of preprocessing. (c) The final segmenting result of region prior based GAC.

3.2 Shape Prior Based GAC

Because of the intrinsic noise of the echocardiographic image and the blur caused by movement of the cardiac valve, it is unavoidable that there would be some segmentation errors in the final result. To get a more accurate contour, we need make full use of the prior shape of heart valve in segmenting process. The prior shape could be a manual outline, acceptable neighboring final result, or statistical contour.

A speed field, which directs to the nearest point of the prior shape, is set nearby the shape. The force F_{shape} drives all the points nearby to the shape. Consider a prior contour C, we define the distance from point X to contour C as ε:

$$\varepsilon(X) = d(X) = \min(|X - X_I|) \qquad X_I \in C \tag{10}$$

Then the speed field produced by prior contour is:

$$F_{shape}(X) = f_s(d)\frac{\nabla\varepsilon}{|\nabla\varepsilon|} \tag{11}$$

F_{shape} directs to the nearest shape point, the magnitude of F_{shape} is controlled by $f_s()$.

Attribute of the prior shape force is very important. There are two kinds of force. One is like the elastic force which would be more powerful far away from prior shape, the other is just like the force in electric field, the closer to the shape, the more powerful it would be. It is expected that F_{shape} could only take effect in the field nearby the prior contour and leave it alone far away from the prior contour. So in our work, the second kind of force is taken. It's supposed that the farthest neighborhood distance be δ:

$$f_s(d) = \begin{cases} k(\delta - d) & d \le \delta \\ 0 & d > \delta \end{cases} \tag{12}$$

The final shape prior based GAC is:

$$\frac{\partial\phi}{\partial t} = u(x)(k + v_0)|\nabla\phi| + \nabla u \cdot \nabla\phi + f_s(d)\frac{\nabla\varepsilon}{|\nabla\varepsilon|} \cdot \nabla\phi \tag{13}$$

The algorithm is demonstrated on synthetic image. We could get a better result guided by prior shape (Fig. 2). Where δ =15 in segmenting evolution.

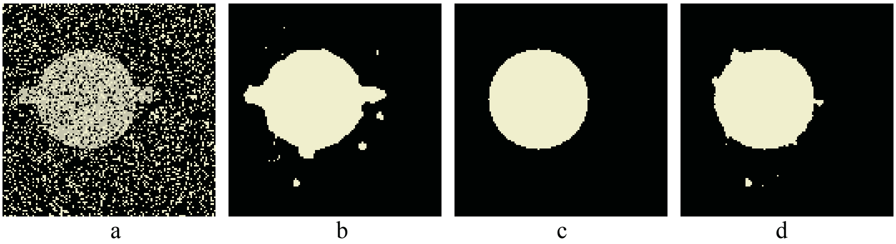


Fig. 2. Circle segmented by shape prior based GAC. (a)Circle stained by a bar and salt & pepper noise. (b) Circle segmented by GAC. (c) Prior circle shape. (d) Circle Segmented by shape prior based GAC.

4 Application and Results

We put the algorithm into practice and efficiently segmented heart valve leaflets from ten 3D echocardiographic sequences, each covering one complete cardiac cycle. The 3D sequences were recorded using the Philips Sonos 750 TTO probe that scanned object rotationally. There are 13-17 frames per cardiac cycle and the angular slice spacing was 3 degrees resulting in 60 images slices in each frames. Therefore, there are about 1000 images in each 3D sequence. The resolution of image was 240*256.

The large quantity of data made it impossible to segment all images manually by traditional method. Region prior based GAC could segment the whole sequence automatically guided by several prior regions without too much parameter choosing or adjusting. And it is neither sensitive to the initial zero level set nor prior region as long as it is just between the valve and cardiac wall. All these make the segmentation procedure more efficient and the segmenting results more precise. Some results of region prior based GAC are shown in Fig. 3. The images in Fig. 3 are at the same scanning position of different scanning time from a 3D valve sequence. To facilitate the display, we cut the valve region out.

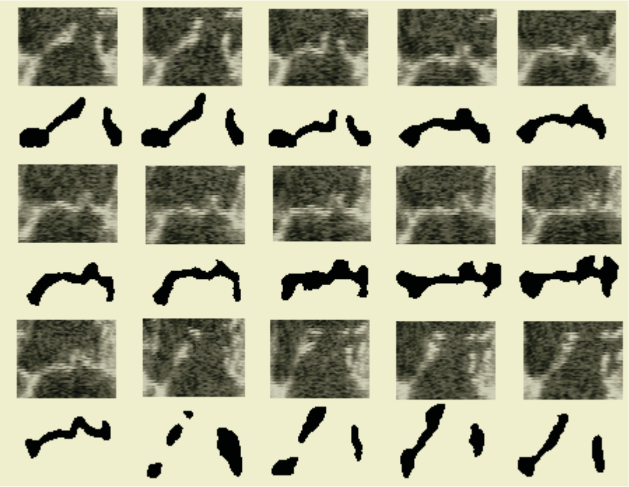


Fig. 3. Valves segmented by region prior based GAC.

Most of segmentation results above could satisfy the needs of 3D reconstruction and diagnosis. But the contamination of noise and motion blur increase the segmentation errors inevitable, such as the eighth, the twelfth, and the thirteenth contour in Fig. 4. We could segment this kind of noise-disturbed images guided by a strong constrain, i.e. shape prior. Two level prior, region and shape, could get a robust and accurate result. The prior shape could either come from an output of neighbor slice, or from a manual outline by expert physicians. And a part of prior shape of stained edge is just enough. We show some results segmented by shape prior based GAC in Fig. 4. The segmentation errors could be restored under the guidance of the shape prior.

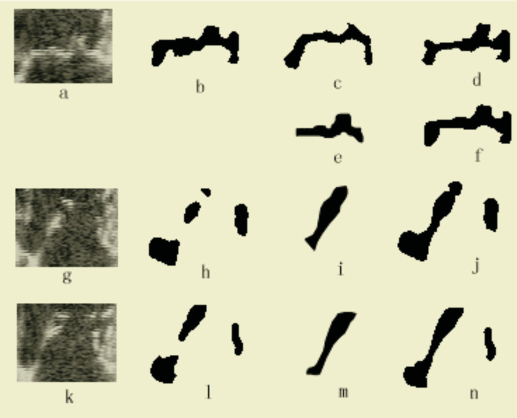


Fig. 4. Cardiac valve segmented by shape prior based GAC. (a), (g), (k) The initial image. (b), (h), (l) Result guided by region prior. (c) Segmentation result of neighbor slice. (d) Result guided by shape prior of c. (e), (i), (m) Manual outline. (f), (j), (n) Result guided by shape prior.

5 Conclusions

The inherent noise, blur and the large quantity of data of echocardiographic sequences make it difficult to segment the valve structures, which hold back the clinical application. We present a new algorithm to incorporate prior knowledge into geodesic active contour. The prior is expressed as a speed field, which directly draws the zero level set to the ideal contour. Region prior constrains the zero level set evolving in certain region and shape prior draws the curve to the ideal contour. The actual application on 3D echocardiographic sequences shows that the algorithm segments the valve structure accurately and reduces the manual procedure greatly resulting in an accelerated segmenting process.

Prior based image segmentation is an active subject at present. More and more prior knowledge should be represented as speed field and embed in image segmentation process in the future work. Guided by prior information, the image segmentation would be more accurate and efficient.

Acknowledgements

This work was partially supported by National Science Research Program of China (No. 2004BA714802) and Shanghai Science and Technology Development Foundation (034119820).

References

1. Kass, M., Witkin, A., and Terzopoulos, D.: Snakes: Active contour models. International Journal of Computer Vision, Vol. 1(1988) 321-331.
2. Daniel, C., Florian, T., Joachim, W. and Christoph, S.: Diffusion Snakes: Introducing Statistical Shape Knowledge into the Mumford-Shah Function. International Journal of Computer Vision, Vol. 50(2002) 295-313.

3. Daniel, C., Timo, K. and Christoph, S.: Shape Statistics in Kernel Space for Variational Image Segmentation. Pattern Recognition, Vol. 36(2003) 1929-1943.
4. Ivana, M., Slawomir, K. and James D.T.: Segmentation and Tracking in Echocardiographic sequences: Active Contour Guided by Optical Flow Estimates. IEEE Trans. on Medical Imaging, Vol. 17(1998) 274-284.
5. Michael, E.L., Grimson, W.E.L. and Olivier, F.: Statistical Shape Influence in Geodesic Active Contours. Computer Vision and Image Understanding, Vol. 1(2000) 316-323
6. Chen, Y., Hemant, D., Tagare, S.T. etc.: Using Prior Shapes in Geometric Active Contours in a Variational Framework. International Journal of Computer Vision, Vol. 50(2002) 315-328.
7. Caselles, V., Kimmel, R. and Sapiro. G.: Geodesic Active Contours. International Journal of Computer Vision, Vol. 22(1997) 61-79.
8. Kichenassamy, A., Kumar, A., Olver, P. etc.: Gradient Flows and Geometric Active Contour Models. In IEEE Int'l Conf. Comp. Vision, (1995) 810-815.
9. Pietro, P. and Jalhandra, M.: Scale-space and Edge Detection Using Anisotropic Diffusion. IEEE Trans. on Pattern Analysis Machine Intelligence, Vol.12(1990) 629-639.
10. Whitaker, R. and Xue, X.: Variable-Conductance, Level-Set Curvature for Image Processing, ICIP, (2001) 142-145.

Bayesian Reconstruction for Transmission Tomography with Scale Hyperparameter Estimation[⋆]

Antonio López[1], Rafael Molina[2], and Aggelos K. Katsaggelos[3]

[1] Universidad de Granada, Departamento de Lenguajes y Sistemas Informáticos,
18071 Granada, Spain
alopez@ugr.es

[2] Universidad de Granada. Departamento de Ciencias de la Computación e I.A.,
18071 Granada, Spain
rms@decsai.ugr.es

[3] Northwestern University. Department of Electrical and Computer Engineering,
Evaston, Illinois 60208-3118
aggk@ece.northwestern.edu

Abstract. In this work we propose a new method to estimate the scale hyperparameter for transmission tomography in Nuclear Medicine image reconstruction problems. Within the Bayesian paradigm, Evidence Analysis and circulant preconditioners are used to obtain the scale hyperparameter. For the prior distribution, we use Generalized Gaussian Markov Random Fields (GGMRF), a nonquadratic function that preserves the edges in the reconstructed image. The experimental results indicate that the proposed method produces satisfactory reconstructions.

1 Introduction

PET (positron emission tomography) and SPECT (single photon emission tomography) are techniques used in Nuclear Medicine to obtain cross sectional images which represent an isotope distribution within the body of a patient [3].

The attenuation or absorption of photons is an important effect in PET and SPECT systems, that produces errors in emission tomography due to decreasing quantitative accuracy of the reconstructed image. The attenuation correction factors (ACFs) are obtained from a transmission scan that determines the tissue structure in a patient. The transmission scan can be either previous (pre-injection measurements) or simultaneous to the emission scan are (post-injection measurements).

Several Bayesian methods have been proposed for transmission tomography, see for instance [4, 10]. These methods use a prior model incorporating the expected structure in the image. These image models depend on parameters known as hyperparameters and an optimal Bayesian reconstruction needs appropriate

[⋆] This work has been partially supported by "Instituto de Salud Carlos III" project FIS G03/185 and by CICYT project TIC2000-1275.

hyperparameter values. Therefore, reliable automatic methods for the selection of the hyperparameters are essential to obtain correct reconstructions. Unfortunately, not much work has been reported on the parameter estimation for transmission tomography, see however [7].

In this paper, we propose an estimation method of the scale hyperparameter that can be used to reconstruct attenuation maps for pre-injection measurements.

The rest of the paper is organized as follows. In section 2 we describe the transmission and image models and the Evidence Analysis within the Bayesian paradigm. Section 3 describes the proposed estimation method. Experimental results are presented in section 4 and section 5 concludes the paper.

2 Hierarchical Bayesian Paradigm and Evidence Analysis

Within the Bayesian paradigm, the reconstruction of the original image X, denoted by $\hat{X}(\theta)$, is selected as:

$$\hat{X}(\theta) = \arg\max_{X} P(Y|X)P(X|\theta) \ , \tag{1}$$

where θ is a hyperparameter vector, $P(X|\theta)$ is the prior distribution, and $P(Y|X)$ models the process to obtain the data Y from the real underlying image X.

The Hierarchical Bayesian paradigm first defines the distributions $P(Y|X)$ and $P(X|\theta)$. Next, a distribution $P(\theta)$ for the hyperparameters is defined and the joint distribution $P(\theta, X, Y)$ is formed. Using the Evidence Analysis, we perform the following steps to estimate the hyperparameter vector and reconstruct the image:

1. $P(\theta, X, Y)$ is integrated over the whole image space X to obtain the distribution $P(\theta, Y)$ and
$$\hat{\theta} = \arg\max_{\theta} P(\theta, Y)P(\theta) \ , \tag{2}$$
is selected as the hyperparameter vector.
2. The original image X is estimated as:
$$\hat{X}(\hat{\theta}) = \arg\max_{X} P(Y|X)P(X|\hat{\theta}) \ . \tag{3}$$

In order to define $P(Y|X)$, we note that in transmission tomography the attenuation is independent of position along the projection line and the observation data and is specified as Poisson distributions (pre-injection measurements):

$$\log P(Y|X) \propto \sum_{i=1}^{M} \{-b_i \exp\{-\sum_{j=1}^{N} A_{i,j}x_j\} + y_i \log(b_i \exp\{-\sum_{j=1}^{N} A_{i,j}x_j\})\} \ , \tag{4}$$

where M is the number of detectors, N the number of pixels, and A the system matrix. $A_{i,j}$ is the intersection length of the projection line i with the area represented by pixel j, x_j represents the attenuation correction factor at pixel

j, b_i is the blank scan and y_i is the number of transmission counts at the ith detector pair in PET or detector in SPECT. We note that in this distribution the random coincidences are ignored. For transmission tomography is known that the $ACFs$ within each tissue of a patient are homogeneous and they present abrupt variations in the transition between tissues. Therefore, nonquadratic prior distributions are extensively used for transmission reconstruction [12].

In this work we use generalized Gaussian Markov random fields ($GGMRF$) as prior models [2]. This distribution has the following form:

$$P(X|\theta)=\frac{1}{Z(\theta)}\exp\{-\frac{1}{p\sigma^p}\sum_{i,j\in\mathcal{N}}U(x_i,x_j,p)\}=\frac{1}{Z(\sigma,p)}\exp\{-\frac{1}{p\sigma^p}\sum_{i,j\in\mathcal{N}}w_{i-j}|x_i-x_j|^p\}, \tag{5}$$

where $\theta=(p,\sigma)$ is the hyperparameters vector, U the energy function and Z the partition function. The elements x_i and x_j are neighbouring pixels and $\mathcal{N}$ is the set of all neighbouring pixel pairs. The scale hyperparameter σ determines the overall smoothness of the reconstruction and p is called the shape parameter. The potential function is convex when $p>1$. Since our energy function is scalable, that is, for all $X\in\mathcal{R}^N$ and $\alpha>0$ we have:

$$U(\alpha X,p)=\alpha^p U(X,p)\ , \tag{6}$$

it follows that the partition function can be expressed as:

$$Z(\sigma,p)=(p\sigma^p)^{N/p}Z(1,p). \tag{7}$$

Equation (7) implies that the partition function is derivable with respect to σ. Hence, a method for the estimation of σ can be obtained.

Following the discussion in [11], we do not estimate the parameter p. Since values of p close to 1 better preserve the formation of edges, this parameter was set equal to 1.1.

3 Scale Hyperparameter Estimation

We now proceed to estimate the scale hyperparameter for transmission reconstruction problems. We assume that $P(\theta)\propto$ const.

In order to solve Eq. (3), we define the following function $M(X,Y|\theta)$:

$$\begin{aligned}M(X,Y|\theta)&=\log P(X|\theta)-\log P(Y|X)\propto-\log Z(\sigma,p)-\tfrac{1}{p\sigma^p}\textstyle\sum_{i,j\in\mathcal{N}}U(x_i,x_j,p)\\&+\textstyle\sum_{i=1}^{M}\{b_i\exp\{-\sum_{j=1}^{N}A_{i,j}x_j\}-y_i\log(b_i\exp\{-\sum_{j=1}^{N}A_{i,j}x_j\})\},\end{aligned} \tag{8}$$

and then we obtain $P(\theta,Y)$ using

$$P(\theta,Y)\propto P(\theta)\int_X\exp\{M(X,Y|\theta)\}dX\ . \tag{9}$$

The integral in Eq. (9) cannot be evaluated analytically and therefore we resort to Gaussian quadrature approximation. Using Taylor series expansion, we expand $M(X, Y|\theta)$ around the MAP estimate of X given θ, $\hat{X}(\theta)$, see Eq. (3). Keeping up to second order terms, we have the following approximation of $P(\theta, Y)$:

$$P(\theta, Y) \propto \exp\left[M(\hat{X}(\theta), Y|\theta)\right] \left|G(\hat{X}(\theta)) + \frac{1}{p\sigma^p} F(\hat{X}(\theta))\right|^{-1/2} , \qquad (10)$$

where the (i, j)th elements of the matrices $G(\hat{X}(\theta))$ and $F(\hat{X}(\theta))$ are given by:

$$G_{i,j}(\hat{X}(\theta)) = \sum_{k=1}^{M} b_k A_{k,i} A_{k,j} \exp\left(-\sum_{l=1}^{N} A_{k,l}\hat{x}_l(\theta)\right) , \qquad (11)$$

$$F_{i,j}(\hat{X}(\theta)) = \begin{cases} \sum_{k\in\mathcal{N}_i} \frac{\partial^2 U(\hat{x}_i(\theta), \hat{x}_k(\theta), p)}{\partial x_i^2} & i = j \\ -\frac{\partial^2 U(\hat{x}_i(\theta), \hat{x}_j(\theta), p)}{\partial x_i \partial x_j} & i \neq j,\ j \in \mathcal{N}_i \\ 0 & \text{otherwise} . \end{cases} \qquad (12)$$

Using Eq. (10) in Eq. (2) we obtain:

$$\sigma^p = \frac{1}{N} \sum_{i,j\in\mathcal{N}} U(\hat{x}_i(\theta), \hat{x}_j(\theta), p) + \frac{1}{2N} trace\left[\left(G(\hat{X}(\theta)) + \frac{1}{p\sigma^p} F(\hat{X}(\theta))\right)^{-1} F(\hat{X}(\theta))\right] . \qquad (13)$$

We have the following iterative procedure to estimate the attenuation map X and the scale hyperparameter σ. At each step k, we proceed as follows:

1. Given a previously obtained attenuation map $\hat{X}_{k-1}$ and a previous estimate of the scale hyperparameter $\hat{\sigma}_{k-1}$, we obtain a new value $\hat{\sigma}_k$ by using Eq. (13).
2. Given $\hat{\sigma}_k$ and $\hat{X}_{k-1}$, a newly estimated attenuation map $\hat{X}_k$, is obtained by iterating once a MAP algorithm.
3. We set $k = k + 1$, and go to step 2 until a stop criterium with respect to σ is satisfied.

We note that the estimation of σ involves the calculation of $(G(\hat{X}(\theta)) + \frac{1}{p\sigma^p} F(\hat{X}(\theta)))^{-1}$. This inversion is a computationally intensive problem. We use preconditioning to approximate these matrices in order to simplify their inversion. Diagonal and circulant preconditioning methods are used to approximate the matrix $(G(\hat{X}(\theta)) + \frac{1}{p\sigma^p} F(\hat{X}(\theta)))$. These preconditioner were previously applied to emission SPECT reconstruction [8, 9].

3.1 Diagonal Preconditioner

By using only the diagonal elements of the matrices (i. e., ignoring the off-diagonal elements), we obtain:

$$\sigma^p = \frac{1}{N} \sum_{i,j\in\mathcal{N}} w_{i-j} \left|\hat{x}_i(\theta) - \hat{x}_j(\theta)\right|^p + \frac{p(p-1)}{2N} \frac{\sum_{k\in\mathcal{N}_j} w_{j-k} \left|\hat{x}_j(\theta) - \hat{x}_k(\theta)\right|^{p-2}}{a_j(\theta)} , \qquad (14)$$

where

$$a_j(\theta) = \sum_{k=1}^{M} b_k A_{kj}^2 \exp\left(- \sum_{l=1}^{N} A_{k,l} \hat{x}_l(\theta) \right) + \frac{p-1}{\sigma^p} \sum_{k \in \mathcal{N}_j} w_{j-k} \left| \hat{x}_j(\theta) - \hat{x}_k(\theta) \right|^{p-2} .$$

We note that when $p < 2$, $|z|^{p-2} = \infty$ for $z = 0$. However, this is not a problem due to the form of Eq. (14) and the presence of $a_j(\theta)$.

3.2 Circulant Preconditioner

Following [8], we express the Hessian matrix of the transmission model, $G(\hat{X}(\theta))$, as:

$$G(\hat{X}(\theta)) = A^t W(\hat{X}(\theta)) A , \tag{15}$$

where $W(\hat{X}(\theta))$ is a diagonal matrix with diagonal entries

$$W_{i,i}(\hat{X}(\theta)) = b_i \exp\left(\sum_{l=1}^{N} A_{i,l} \hat{x}_l(\theta) \right) . \tag{16}$$

We then apply the approximation introduced in [6] to the Fisher information term, which produces:

$$A^t W(\hat{X}(\theta)) A \approx D_{(g)}(\hat{X}(\theta)) A^t A D_{(g)}(\hat{X}(\theta)) , \tag{17}$$

where $D_{(g)}(\hat{X}(\theta))$ is a diagonal matrix with diagonal entries

$$D_{(g)j,j}(\hat{X}(\theta)) = \sqrt{\frac{\sum_{i=1}^{M} A_{i,j}^2 W_{i,i}(\hat{X}(\theta))}{\sum_{i=1}^{M} A_{i,j}^2}} . \tag{18}$$

The system matrix A is shift variant, and so the product $A^t A$ is approximated by a block-circulant matrix G_c. The kernel of G_c is obtained as follows:

1. First we calculate $A^t A \epsilon_j$, where ϵ_j represents a unit vector centered with respect to the image.
2. Then, the kernel obtained in the previous step, is approximated by a shift invariant symmetric blurring function.

We also apply the circulant approximation proposed in [6] to the Hessian matrix of our prior model, $F(\hat{X}(\theta))$, :

$$F(\hat{X}(\theta)) \approx D_{(f)}(\hat{X}(\theta))(I - \phi C) D_{(f)}(\hat{X}(\theta)) , \tag{19}$$

where the diagonal elements of $D_{(f)}(\hat{X}(\theta))$ have the form:

$$D_{(f)j,j}(\hat{X}(\theta)) = \sqrt{\sum_{k \in \mathcal{N}_j} \frac{\partial^2 U(\hat{x}_j(\theta), \hat{x}_k(\theta), p)}{\partial x_j^2}} , \tag{20}$$

and $(I-\phi C)^{-1}$ is a covariance matrix, with $C_{i,j}=1$ for two neighbouring pixels and for the 8-point neighborhood system we are using, ϕ is just less than 1/8.

The matrices $D_{(g)}(\hat{X}(\theta))$ and $D_{(f)}(\hat{X}(\theta))$ in Eq. (18) and Eq. (20) are approximated by constant diagonal matrices $\sqrt{\alpha(\hat{X}(\theta))}I$ and $\sqrt{\beta(\hat{X}(\theta))}I$ respectively, where:

$$\sqrt{\alpha(\hat{X}(\theta))}=\frac{1}{N}\sum_{j=1}^{N}D_{(g)j,j}(\hat{X}(\theta)) \text{ and } \sqrt{\beta(\hat{X}(\theta))}=\frac{1}{N}\sum_{j=1}^{N}D_{(f)j,j}(\hat{X}(\theta)) \,. \tag{21}$$

Using the above approximations we now have,

$$trace\left[\left(G+\frac{1}{p\sigma^p}F\right)^{-1}F\right]\approx trace\left[\left(\alpha G_c+\frac{1}{p\sigma^p}\beta(I-\phi C)\right)^{-1}\beta(I-\phi C)\right] \tag{22}$$

where we have removed the dependency of the matrices on $\hat{X}(\theta)$ for simplicity. Finally, using the DFT we obtain the following expression:

$$trace\left[\left(G+\frac{1}{p\sigma^p}F\right)^{-1}F\right]\approx\sum_{i=1}^{N}\frac{\Lambda_i}{\frac{\alpha}{\beta}\Gamma_i+\frac{1}{p\sigma^p}\Lambda_i}\,, \tag{23}$$

where Γ_i and Λ_i are the (i)th elements of the DFT of G_c and $(I-\phi C)$, respectively.

When $p<2$, the potential function has non bounded second derivative for differences $x_i-x_j=0$. This is a problem for the circulant approach, since, it requires the potential function to be twice-differentiable with respect to X and with bounded second derivatives. In order to overcome this problem, we have used the approximation suggested in [1] for GGMRF prior models:

$$U(X,p)=\sum_{i,j\in\mathcal{N}}w_{i-j}|x_i-x_j|^p\approx\sum_{i,j\in\mathcal{N}}w_{i-j}\left(\left(|x_i-x_j|^2+\delta\right)^{p/2}-\delta^{p/2}\right)\,, \tag{24}$$

where $\delta>0$ is a stabilization constant. Therefore, $\beta(\hat{X}(\theta))$ is obtained by using this approximation of $U(X,p)$.

4 Experimental Results

The proposed reconstruction method was applied to real PET transmission data. These data are available in [5] and represent a PET scan of an anthropomorphic torso phantom (Data Spectrum Corporation). The blank and transmission scan have 192 angles and 160 radial bins and the size of the reconstructed attenuation map is 128×128 pixels.

We denote by DP the estimation method using diagonal preconditioner, described in section 3.1, and by CP the estimation method using the circulant approach, described in section 3.2. The image update in step 2 of the algorithm in section 3 was obtained using the MAP algorithm proposed in [4].

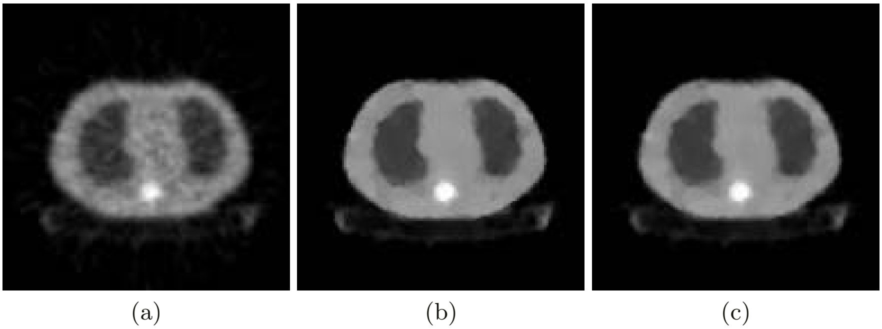


Fig. 1. (a) FBP reconstruction. (b) DP reconstruction. (c) CP reconstruction.

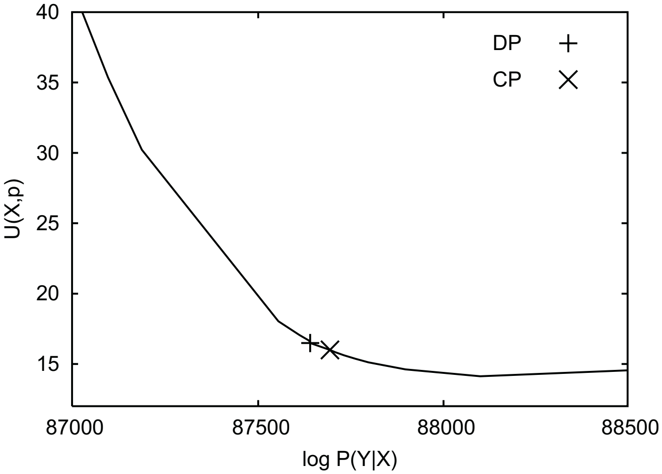


Fig. 2. L-curve. The obtained values of $U(X, p)$ and $\log P(Y|X)$ with DP (+) and CP (×) methods have been marked in the curve.

The estimations provided by the methods are $\hat{\sigma}^{1.1} = 0.0069$ (DP), and $\hat{\sigma}^{1.1} = 0.0058$ (CP). The reconstructions obtained with these values are show in Fig. 1(b) and Fig.1 (c). For comparison, a filtered back projection (FBP) reconstruction is shown in Fig. 1(a). We observe that both diagonal and circulant approaches provide a better reconstructions, having less noise that the FBP reconstruction.

Using as stopping criterion $\left|(\hat{\sigma}^{1.1})^k - (\hat{\sigma}^{1.1})^{k-1}\right| / (\hat{\sigma}^{1.1})^k \leq 0.0001$ the number of iterations needed was: 71 (DP) and 38 (CP).

In order to quantitatively evaluate the obtained reconstructions, we have computed the L-curve which represents the value of the function energy $U(X, p)$ versus to the $\log P(Y|X)$ for the MAP reconstruction, obtained for a range of scale hyperparameter $\sigma^{1.1}$ values. Heuristically, the maximum curvature point of the L-curve corresponds to a good balance between fidelity to the observation data and smoothness of the reconstruction. The obtained values of $U(X, p)$ and $\log P(Y|X)$ with the DP and CP methods have been marked in the L-curve. They lie within the zone of the maximum curvature of the L-curve (see Fig. 2).

5 Conclusions

In this paper we have presented a method to estimate the scale hyperparameter σ to reconstruct attenuation maps in the transmission tomography context. The application of preconditioning methods to estimate this unknown hyperparameter has been described. Both with the diagonal and circulant preconditioners, we have found that the estimated values of the scale hyperparameter are close to the value considered as optimun (σ at the maximum curvature point of the L-curve). The circulant approach exhibits a better convergence rate than the diagonal approach.

Although we have used GGMRF priors, the proposed method is suitable for other convex priors as well, if they have a partition function for which its derivative with respect to the scale hyperparameter can be obtained.

References

1. Belge, M., Kilmer, M.E., Miller, E.L.: Wavelet Domain Image Restoration with Adaptative Edge-Preserving Regularization. IEEE Tr. Im. Proc. **9** (2000) 597–608
2. Bouman, C.A., Sauer, K.: A Generalized Gaussian Image Model for Edge-Preserving Map Estimation. IEEE Tr. Im. Proc. **2** (1993) 296–310
3. Z.H. Cho, J.P. Jones, and M. Singh, Foundation of medical imaging, John Wiley and Sons, 1993.
4. Erdogan, H., Fessler, J.A.: Monotonic Algorithms for Transmission Tomography. IEEE Tr. Me. Im. **18** (1999) 801–14
5. Fessler, J.A.: ASPIRE (A Sparse Precomputed Iterative Reconstruction Library). http://www.eecs.umich.edu/˜fessler/aspire/index.html
6. Fessler, J.A., Booth, S.D.: Conjugate-Gradient Preconditioning Methods for Shift Variant PET Image Reconstruction. IEEE Tr. Im. Proc. **8** (1999) 688–699
7. Hsiao, I., Rangarajan, A, Gindi, G.: Joint MAP Bayesian Tomographic Reconstruction with a Gamma-mixture Prior. IEEE Tr. Im. Proc. **11** (2002) 1466–1477
8. López, A., Molina, R., Katsaggelos, A.K.: Scale Hyperparameter Estimation for GGMRF Prior Models with Application to SPECT Images. Proc. of IEEE Int. Conf. on Digital Signal Processing, Vol. 2 (2002) 521–524
9. López, A., Molina, R., Katsaggelos, A.K., Rodriguez, A., López J.M., Llamas, J.M.: Parameter Estimation in Bayesian Reconstruction of SPECT Images: An Aid in Nuclear Medicine Diagnosis. Int. J. Imaging Syst. Technol. **14** (2004) 21–27
10. Mumcuoglu, E.U., Leahy, R., Cherry, S.R., Zhou, Z.: Fast Gradient-based Methods for Bayesian Reconstruction of Transmission and Emission PET Images. IEEE Tr. Me. Im. **13** (1993) 687–701
11. Saquib, S.S.: Edge-preserving Models and Efficient Algorithms for ill-osed Inverse Problems in Image Processing. Ph.D. thesis, University of Purdue, 1997.
12. Yu, D.F., Fessler, J.A.: Edge-preserving Tomographic Reconstruction with Nonlocal Regularization. IEEE Tr. Me. Im. **21** (2002) 159–73

Automatic Segmentation and Registration of Lung Surfaces in Temporal Chest CT Scans

Helen Hong[1], Jeongjin Lee[2], Yeni Yim[2], and Yeong Gil Shin[2]

[1] School of Computer Science and Engineering,
BK21: Information Technology,
Seoul National University,
San 56-1 Shinlim 9-dong Kwanak-gu, Seoul 151-742, Korea
hlhong@cse.snu.ac.kr
[2] School of Computer Science and Engineering,
Seoul National University,
San 56-1 Shinlim-dong Kwanak-gu, Seoul 171-542, Korea
{jjlee,shine,yshin}@cglab.snu.ac.kr

Abstract. We propose an automatic segmentation and registration method for matching lung surfaces of temporal CT scans. Our method consists of three steps. First, an automatic segmentation is used for accurately identifying lung surfaces. Second, initial registration using an optimal cube is performed for correcting the gross translational mismatch. Third, the initial alignment is step by step refined by the iterative surface registration. For the fast and robust convergence of the distance measure to the optimal value, a 3D distance map is generated by the narrow band distance propagation. Experimental results show that our segmentation and registration method extracts accurate lung surfaces and aligns them much faster than conventional ones using a distance measure.

1 Introduction

Chest computed tomography (CT) is a well-establishing means of diagnosing pulmonary metastasis of oncology patients and evaluating lung disease progression and regression during treatment. With ever-improving resolution and availability of CT scanners, chest CT is being used as a successful screening method for identifying early lung cancer. Most patients undergoing screening for lung cancer have non-calcified nodules or small nodules less than 5mm in diameter. Such small nodules are often very difficult to characterize, because they are commonly benign or early malignancy. To estimate the probability of malignancy, we need to investigate changes of small nodules through time-interval screening processes. For the follow-up study of pulmonary nodules, we suggest an automatic segmentation and registration method of corresponding lung surfaces in temporal chest CT scans.

Several methods have been suggested for the matching of lung surfaces in temporal chest CT scans. In Betke et al. [1], anatomical landmarks such as the sternum, vertebrae, and tracheal centroids are used for initial global registration. Then the initial surface alignment is refined by an iterative closest point (ICP) process. Most part of the computation time for the ICP process is to find the point correspondences of lung surfaces obtained from two time interval CT scans. Hong et al. [2] proposed an effi-

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 463–470, 2005.

cient multilevel method for surface registration to cope with the problem of Betke [1]. The multilevel method first reduces the original number of points and aligns them using the ICP algorithm. In addition, they proposed a midpoint approach to define the point correspondence instead of using the point with the smallest Euclidean distance as in the original ICP algorithm. However the midpoint approach has a tradeoff between accuracy and efficiency, because additional processing time is needed to find the second closest point and to compute the midpoint. Mullaly et al. [3] developed a multi-criterion nodule segmentation and registration method that improves the identification of corresponding nodules in temporal chest CT scans. The method requires additional nodule segmentation process and measures for multi-criterion. Gurcan et al. [4] developed an automated global matching of temporal thoracic helical CT scans. This method uses three-dimensional anatomical information such as ribs without requiring any anatomical landmark identification or organ segmentation. But it would be difficult to align correctly since the method uses only limited information obtained by Maximum Intensity Projection (MIP) images of two time-interval CT scans.

Current approaches still need more progress in computational efficiency and accuracy for investigating changes of pulmonary nodules in temporal chest CT scans. In this paper, we propose an automatic segmentation and registration method which provides more efficiency and robustness for matching time-interval lung surfaces. Our method consists of three steps. First, lungs are extracted from chest CT scans by the automatic segmentation method. Second, the gross translational mismatch is corrected by the optimal cube registration. This initial registration does not require extracting any anatomical landmarks. Third, the initial alignment is step by step refined by the iterative surface registration. To evaluate the distance measure between surface boundary points, a 3D distance map is generated by the narrow band distance propagation, which drives fast and robust convergence to the optimum value. Experimental results show that our segmentation and registration method extracts accurate lung surfaces and aligns them much faster than conventional ones using a 3D distance map. Accurate and fast result of our method would be more useful for clinical application of lung cancer screening.

The organization of the paper is as follows. In Section 2, we discuss how to extract the lungs from other organs in chest CT scans and how to correct the geometrical mismatch. In Section 3, experimental results show how our segmentation method accurately extracts the lungs and how our registration method rapidly aligns lung surfaces of two time CT scans. This paper is concluded with brief discussion of the results in Section 4.

2 Automatic Lung Segmentation and Registration

For the registration of the current CT scan, called target volume, with the previous one, called template volume, we apply the pipeline shown in Fig. 1. Since our method is applied to the lung cancer screening, we assume that each CT scan is acquired at the maximal inspiration and the dataset includes the thorax from the trachea to below the diaphragm. Based on this assumption and experience, we found that rigid transformation would be sufficient for the registration of temporal chest CT scans.

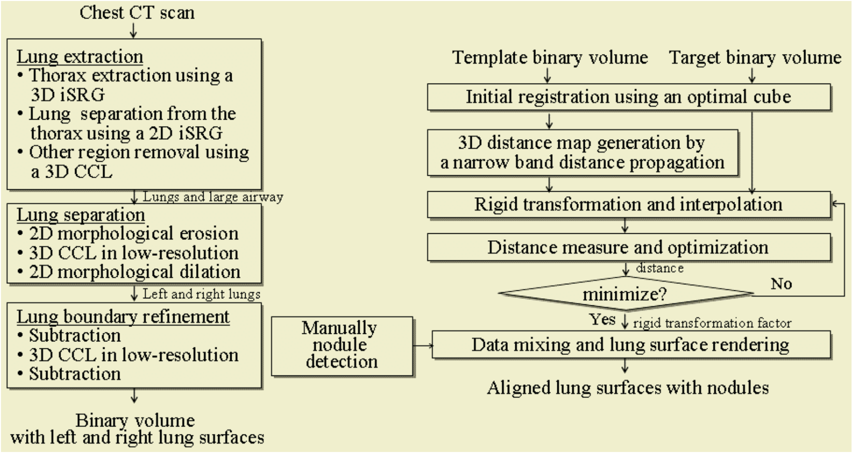

Chest CT scan
Lung extraction
• Thorax extraction using a 3D iSRG
• Lung separation from the thorax using a 2D iSRG
• Other region removal using a 3D CCL
Lungs and large airway
Lung separation
• 2D morphological erosion
• 3D CCL in low-resolution
• 2D morphological dilation
Left and right lungs
Lung boundary refinement
• Subtraction
• 3D CCL in low-resolution
• Subtraction
Binary volume with left and right lung surfaces
Manually nodule detection
Template binary volume
Target binary volume
Initial registration using an optimal cube
3D distance map generation by a narrow band distance propagation
Rigid transformation and interpolation
Distance measure and optimization
distance
minimize?
No
Yes
rigid transformation factor
Data mixing and lung surface rendering
Aligned lung surfaces with nodules

Fig. 1. The pipeline of the automatic lung segmentation and registration.

2.1 Automatic Lung Segmentation Using a Hybrid Approach

A precursor to the whole process for matching lung surfaces in two-time interval chest CT scans is lung segmentation. In particular, with the introduction of multislice spiral CT scanners bringing a large number of volumetric studies of the lung, it is critical to develop an efficient method that requires minimal or no human interaction for segmenting the precise lung boundary. In this section, we describe a fully automatic segmentation method for accurately identifying lungs in chest CT scans. The method consists of three steps: 1) The extraction step to identify the lungs, 2) the separation step to delineate the trachea and large airways from the lungs, 3) the refinement step to obtain satisfactory lung region borders.

The goal of the lung extraction step is to separate voxels of lung tissue from the surrounding anatomy. Generally, thresholding or 3D region growing is used to identify lungs [5]. Since high-density vessels in a lung are identified as lung voxels during the thresholding or 3D region growing process, the three-dimensional lung regions may contain unwanted interior cavities. To fill the lung regions and eliminate these interior cavities, the additional processing such as hole filling is required. To eliminate this process, we apply 3D inverse seeded region growing (iSRG) for extracting the thorax from the surrounding anatomy followed by 2D iSRG for identifying the lungs in the thorax. For 3D and 2D iSRG, a seed point is automatically selected in the outside point of the thorax and of the lungs, respectively. The 3D connected component labeling is then applied to remove voxels of air which surrounds the body, lungs, and other low-density regions within the volume. To reduce the memory use and computation time, the 3D connected component labeling is performed in low-resolution volume. Since the intensities of the trachea and large airways are similar to those of the lungs, the lungs resulting from the extraction step still contain the trachea and large airways. The lung separation step removes the trachea and left and right mainstem bronchi. In this step, at first 2D morphological erosion is repeatedly applied to the lungs. 2D connected component labeling in low-resolution is then used to retain the only two largest components in the volume. Finally, a 2D morphological dilation is proceeded iteratively to restore the approximated original boundary shape without reconnecting the trachea and large airways. After the lung separation step, some region of lung boundaries may be eroded. A lung refinement step gives an accurate and smoothed lung region borders. Fig. 2 shows the process of the lung refinement step. Fig. 2(a) and (b) is the results of the lung extraction step and the lung separation step, respectively. After subtracting Fig. 2(b) from Fig. 2(a), the largest connected component is detected by the 3D connected component labeling as in Fig. 2(d). The smoothed lung region borders are obtained by subtracting Fig. 2(d) from Fig. 2(a).

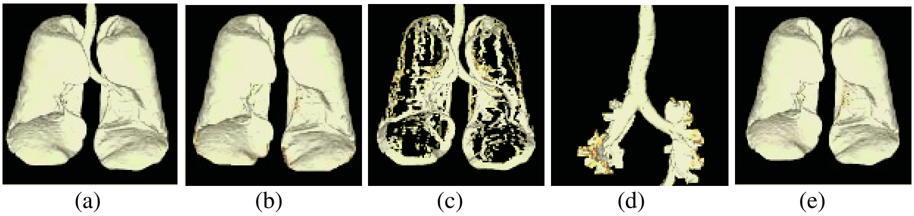


Fig. 2. The process of the lung refinement step (a) the result of the lung extraction step (b) the result of the lung separation step (c) subtraction of b from a (d) the largest component of c (e) subtraction of d from a.

2.2 Initial Registration Using an Optimal Cube

According to the imaging protocol and the patient's respiration and posture, the position of lung surfaces between template and target volume can be quite different. For the efficient registration of such volumes, an initial gross correction method is usually applied. Several landmark-based registrations have been used for the initial gross correction. To achieve the initial alignment of lung surfaces, these landmark-based registrations require the detection of landmarks and point-to-point registration of corresponding landmarks. These processes much degrade the performance of the whole system.

To minimize the computation time and maximize the effectiveness of initial registration, we propose a simple method of global alignment using the circumscribed boundary of lung surfaces. A bounding volume, which includes left and right lung surfaces, is enlarged by a predefined distance *d* to be an optimal cube. The initial registration of two volumes is accomplished by aligning the centers of optimal cubes. The processing time of initial registration using an optimal cube is dramatically reduced since it does not require any anatomical landmark detection. In addition, our method leads to robust convergence to the optimal value since the search space is limited near an optimal value.

2.3 Iterative Surface Registration Using a 3D Distance Map

In a surface registration algorithm, the calculation of distance from a surface boundary to a certain point can be done using a preprocessed distance map based on chamfer matching. Chamfer matching reduces the generation time of a distance map by an approximated distance transformation compared to a Euclidean distance transformation. However, the computation time of distance is still expensive by the two-step distance transformation of forward and backward masks. In particular, when the initial alignment almost corrects the gross translational mismatch, the generation of a 3D distance map of whole volume is unnecessary. From this observation, we propose a narrow band distance propagation for the efficient generation of a 3D distance map.

To generate a 3D distance map, we approximate the global distance computation with repeated propagation of local distances within a small neighborhood. To approximate Euclidean distances, we consider 26-neighbor relations for a 1-distance propagation as shown in Eq.(1). The distance value tells how far it is apart from a surface boundary point. The narrow band distance propagation shown in Fig. 3 is applied to surface boundary points only in the template volume. We can generate a 3D distance map very fast since pixels are propagated only in the direction of increasing distances to the maximum neighborhood.

$$DP(i) = \min(\min_{j \in 26-neighbors(i)} (DP(j)+1), DP(i)) \tag{1}$$

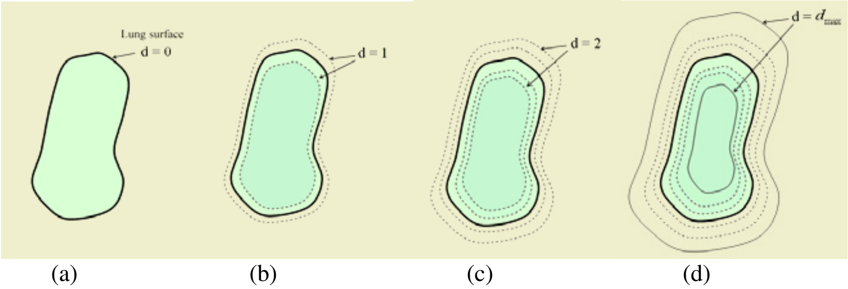


Fig. 3. The generation of the 3D distance map using a narrow band distance propagation (a) lung surface (b) distance 1 propagation (c) distance 2 propagation (d) distance d_{max} propagation.

The distance measure in Eq. (2) is used to determine the degree of resemblance of surface boundaries of template and target volume. The average of absolute distance difference (*AADD*) reaches the minimum when surface boundary points of template and target volumes are aligned correctly. Since the search space of our distance measure is limited to the surrounding lung surface boundaries, the Powell's method is sufficient for evaluating *AADD* instead of using a more powerful optimization algorithm such as simulated annealing.

$$AADD = \frac{1}{N_C} \sum_{i=0}^{N_C - 1} \left| D_{\text{template}}(i) - D_{\text{target}}(i) \right| \tag{2}$$

where $D_{\text{target}}(i)$ is the distance value of target volume and $D_{\text{template}}(i)$ is the distance value of the 3D distance map of template volume. We assume that $D_{\text{target}}(i)$ are all set to 0. N_C is the total number of surface boundary points in target volume.

3 Experimental Results

All our implementation and test have been performed on an Intel Pentium IV PC containing 2.4 GHz CPU and 1.0 GB of main memory. Our method has been applied to ten pairs of successive chest CT scans whose properties are described in Table 1. The performance of our method is evaluated with the aspects of visual inspection and accuracy.

Table 1. Image conditions of experimental datasets

Subject #		Volume size (mm)	Pixel size (mm)	Slice thickness (mm)	# of nodules	Lung volume (cc)	Subject #		Volume size (mm)	Pixel size (mm)	Slice thickness (mm)	# of nodules	Lung volume (cc)
1	Template	512x512x358	0.64x0.64	2.0	1	4469	6	Template	512x512x60	0.57x0.57	5.0	1	3291
	Target	512x512x316	0.66x0.66	2.0		4589		Target	512x512x54	0.63x0.63	5.0		3161
2	Template	512x512x270	0.57x0.57	2.0	1	3229	7	Template	512x512x37	0.57x0.57	8.0	3	3413
	Target	512x512x270	0.55x0.55	2.0		3434		Target	512x512x41	0.52x0.52	8.0		2901
3	Template	512x512x407	0.62x0.62	2.0	1	5773	8	Template	512x512x62	0.59x0.59	5.0	2	4454
	Target	512x512x454	0.64x0.64	2.0		6135		Target	512x512x57	0.65x0.65	5.0		4505
4	Template	512x512x446	0.55x0.55	2.0	2	3102	9	Template	512x512x61	0.62x0.62	5.0	2	3032
	Target	512x512x379	0.54x0.54	2.0		2526		Target	512x512x62	0.71x0.71	5.0		3182
5	Template	512x512x301	0.60x0.60	2.0	8	3287	10	Template	512x512x57	0.63x0.63	5.0	4	3984
	Target	512x512x311	0.51x0.51	2.0		3532		Target	512x512x68	0.62x0.62	5.0		3755

Fig. 4 shows the results of automatic lung segmentation. We can see that lungs with large curvature or complicated shapes are accurately extracted from the chest CT scans.

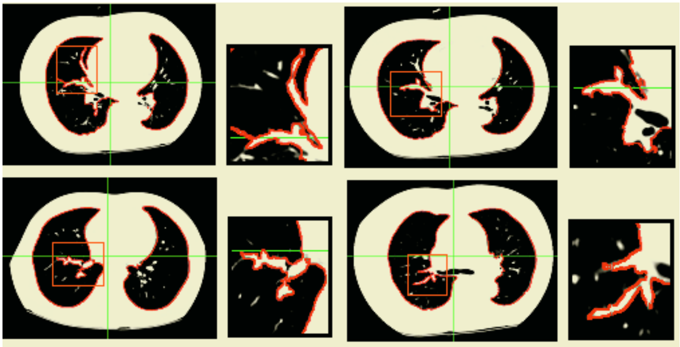
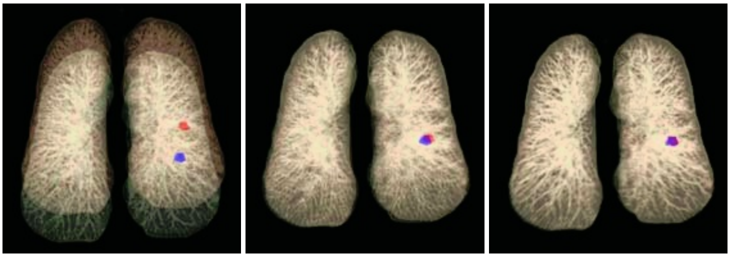


Fig. 4. The results of the automatic lung segmentation.

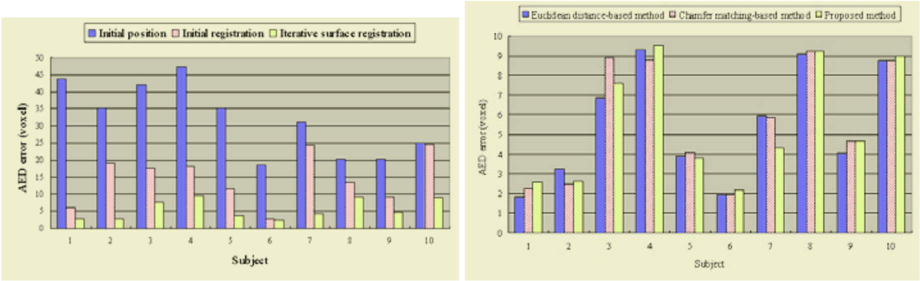
Fig. 5 shows the results of automatic registration of lung surfaces of subject 1. Fig. 5(b) shows the effectiveness of the optimal cube for initial registration. The positional difference between lung surfaces of template and target volumes shown in Fig. 5(a) is much reduced as shown in Fig. 5(b) by the optimal cube registration. This initial alignment is further refined by the iterative surface registration until lung surfaces of template and target volumes are aligned exactly like Fig. 5(c). From the exact matching of lung surfaces, pulmonary nodule correspondences for each subject with nodules in target volume (light sphere) and nodules transformed from template volume into target volume (dark sphere) are shown.



(a) (b) (c)

Fig. 5. The results of the automatic registration of lung surfaces in subject 1 (a) initial position (b) after initial registration (c) after iterative surface registration.

In Fig. 6, the results of pulmonary nodule alignment of ten patients are reported on a per-center-of-mass point basis using the average Euclidean distance (AED) error between corresponding nodules of template and target volumes. In Fig. 6(a), the AED error of most subjects is significantly reduced by the initial registration. In subject 10, the AED error reduction in the initial registration is negligible since there is rotational difference with very little translational difference. In Fig. 6(b), the average value of AED error of the the Euclidean distance-based registration (Method 1), chamfer matching-based registration (Method 2), and our method are 5.48 voxels, 5.69 voxels and 5.55 voxels, respectively. Our method gives similar AED error to Method 1 but more accurate than Method 2.



(a) (b)

Fig. 6. The accuracy evaluation of corresponding nodules using the AED error per subject.

The average of the total processing time of Method 1, Method 2, and our method is 155.2 sec, 76.0 sec and 44.5 sec, respectively. The total processing time of our method is dramatically reduced by the initial registration compared to Method 1 and Method 2. Moreover, our 3D distance map generation time is much faster than that of Method 1 and Method 2.

4 Conclusion

We have developed an accurate and fast method for matching lung surfaces of temporal chest CT scans. Our automatic segmentation using a hybrid approach extracts accurate lung surfaces. Using the optimal cube registration, the initial gross correction of lung surfaces can be done much fast and effective without detecting any anatomical landmarks. In the subsequent iterative surface registration, our distance measure using a 3D distance map generated by the narrow band distance propagation allows rapid and robust convergence to the optimal value. Ten pairs of successive chest CT scans have been used for the performance evaluation with the aspects of visual inspection and accuracy. Our segmentation method gives an accurate lung boundary. In particular, lungs with large curvature or complicated shapes are accurately extracted. Our registration method gives similar accuracy to the Euclidean distance-based registration and much faster than the conventional ones using a 3D distance map. Accurate and fast result of proposed method can be successfully used for investigating temporal changes of pulmonary nodules in lung cancer screening.

References

1. Betke, M., Hong, H., Thomas, D., Prince, C., Ko, J.P., Landmark Detection in the Chest and Registration of Lung Surfaces with an Application to Nodule Registration, Medical Image Analysis, Vol. 7 (2004) 265-281
2. Hong, H., Betke, M., Teng, S., Multilevel 3D Registration of Lung Surfaces in Computed Tomography Scans – Preliminary Experience, Proc. Of International Conference on Diagnostic Imaging and Analysis (ICDIA) (2002) 90-95
3. Mullaly, W., Betke, M., Hong, H., Wang, J., Mann, K., Ko, J.P., Multi-criterion 3D Segmentation and Registration of Pulmonary Nodules on CT: a Preliminary Investigation, Proc. of the International Conference on Diagnostic Imaging and Analysis (ICDIA) (2002) 176-181
4. Gurcan, M.N., Hardie, R.C., Rogers, S.K., Dozer, D.E., Allen, B.H., Hoffmeister, J.W., Automated Global Matching of Temporal Thoracic Helical CT Studies: Feasibility Study, Proc. of International Congress Series, Vol. 1256 (2003) 1031-1036
5. Hu, S., Hoffman, E.A., Automatic Lung Segmentation for Accurate Quantitation of Volumetric X-Ray CT Images, IEEE Trans. on Medical Imaging, Vol. 20, No. 6 (2001) 490-498

Breast Segmentation with Pectoral Muscle Suppression on Digital Mammograms

David Raba, Arnau Oliver, Joan Martí, Marta Peracaula, and Joan Espunya

Robotics and Computer Vision Group, University of Girona, Av. Santalo s/n
Ed. p-IV, 17071 Girona, Spain
{draba,aoliver,joanm,martapb,jespunya}@eia.udg.es
http://vicorob.udg.es

Abstract. Previous works on breast tissue identification and abnormalities detection notice that the feature extraction process is affected if the region processed is not well focused. Thereby, it is important to split the mammogram into interesting regions to achieve optimal breast parenchyma measurements, breast registration or to put into focus a technique when we search for abnormalities. In this paper, we review most of the relevant work that has been presented from 80's to nowadays. Secondly, an automated technique for segmenting a digital mammogram into breast region and background, with pectoral muscle suppression is presented.

1 Introduction

Worldwide, more than 700,000 women die of breast cancer annually and it is estimated that eight to twelve percent of women will develop breast cancer in their lifetime.

Every effort directed to improve early detection is needed. Therefore, many computer vision techniques applied to analysis of digital mammograms have been proposed. Most of them require an initial processing step that splits the image into interesting areas, such as the breast region, background and patient markings. For example, it is well known that information derived from mammographic parenchyma patterns provides one of the most robust signs of risk of developing breast cancer. The largest breast region to be processed, the more accurate the classification of tissue will be. Moreover, the segmentation method should be robust enough to handle a wide range of mammographic images obtained from different image acquisition systems.

This work is part of a larger project called HRIMAC based on the analysis of mammographic images following two different approaches: 1)A Computed Aided Detection platform, which processes the mammograms as a second reader looking for abnormalities using BI-RADS [1] classification, and 2)A featured Computer Aided Diagnosis, which works as a Content Based Image Retrieval System (CBIR). We provide a case with mammogram and personal data and the system retrieves a set of similar cases from the database. This result tries to be a new information source to support radiologist diagnose. In both features,

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 471–478, 2005.

the automatic breast segmentation into background and breast region without artifacts (directly exposed area, the patient identification information and lead markers), is a key objective to provide useful data to the computerized analysis.

In this paper, we will present some of these techniques following this classification: Histogram, Gradient, Polynomial Modelling and Classifier approaches. In section 2.1, we propose an automated method to segment the digital mammogram into breast region and background with a new pectoral muscle suppression technique. Finally experimental and summary conclusions will be presented.

2 Works on Breast Region Segmentation

The breast gross-segmentation have been treated widely. Table 1 shows the tendencies and distribution of methods from the firsts works to recent approaches.

Table 1. Classification of breast gross-segmentation proposals.

Methods	1980's	1990's	2000's
Histogram	Hoyer79 [2]	Lau91 [3] Yin91 [5] Bick95 [6] Byng96 [7] Hein98 [8]	Masek00 [4]
Gradient	Semmlow80 [9]	Méndez96 [10] Abdel-Mottaleb96 [12] Morton96 [13] Karssemeijer97 [14]	Zhou01 [11]
Polynomial Modelling		Stomatakis94 [15] Chandrasekhar96 [16] Goodsitt98 [17]	
Active Contours		Ojala99 [18]	Ferrari00 [19] McLoughlin00 [20] Wirth04 [21]
Classifiers		Lou91 [22]	Saha01 [23] Rickard03 [24] Wirth04 [25] Tromans04 [26]

- ***Histogram based techniques.*** Probably one of the first attempts to separate the breast region was presented by Hoyer *et al* [2] and it was done using simple histogram thresholding. The works of Lau *et al* [3], as well as Yin *et al* [5], and Byng *et al* [7] used a simple thresholding to segment the breast from the background. The work of Bick *et al* [6] presents a combination of local thresholding, region growing and morphological filtering. Hein *et al* [8] propose their own global histogram thresholding, while Masek *et al* [4] proposed a local thresholding method.

- ***Gradient based techniques.*** Breast region extraction techniques based on gradient have long been in use, since the early work of Semmlow *et al* [9], who by means of spatial filters and a Sobel edge detector obtains the breast boundary. Similarly, Méndez *et al* [10] use a two-level histogram threshold to obtain the breast region and oriented upwards, the region is then divided into three parts to track the boundary using the gradient. An evaluation of the quality of the segmentation is provided using the "accurate" or "near accurate" labels. They compare successfully their results to the work presented by Yin *et al* [5]. The work presented by Karssemeijer *et al* [14] takes advantage of a multiresolution scheme, processing in low-res and extrapolating the result. Using a global thresholding technique they obtain a preliminary region, which is processed using a 3x3 Sobel operator, and the pectoral muscle position is estimated via Hough transform. Abdel-Mottaleb *et al* [12] provide an scheme based on different thresholds to find the breast edge. Using the gradient of two images and its union they obtain a possible breast contour. They found the boundary in 98% of the 500 images tested. The segmentation presented by Morton *et al* [13] was another gradient based method. After subtracting the background via an initial threshold, an edge was found by a line-by-line gradient analysis. Zhou *et al* [11] presented an improvement of this last approach.
- ***Polynomial Modelling based techniques.*** An early method proposed by Stomatakis *et al* [15] was not a strict polynomial modelling. By means of an image preprocessing technique to enhance the response of non-dark pixels, a noise reduction process and a histogram threshold, they obtain the breast region, but the boundary is smoothed using Cubic B-splines and samples at fixed pixel intervals are extracted. Then a smooth curve is generated through cubic polynomial calculations. One of the firsts, effective and real polynomial modelling was presented by Chandrasekhar and Attikiouzel [16]. An initial threshold is used to approximate the breast region. Their method provides around 94% acceptable results from 300 images from MIAS [27] mammogram database. A quadratic/cubic polynomial fitting method was proposed by Goodsitt *et al* [17] which is fitted by translation and rotation the axes.
- ***Active Contours based techniques.*** One of the firsts applications of the active contours on breast segmentation was presented by Ojala *et al* [18]. McLoughlin *et al* [20]. They apply a global threshold to obtain an initial result. They statistically model the breast with the pixels inside the region and a snake algorithm is applied to obtain the final boundary. On the other hand, Ferrari *et al* [19] propose a method that firstly enhances the image with a logarithmic transformation, and then an iterative technique (as the Lloyd-Max least-squares) is applied to find and optimal threshold. Finally, they use a B-Spline to approximate the boundary. Recently, Wirth *et al* [21] propose an active contour to segment the breast. The method obtains two preliminary regions using a convolution matrix to enhance the edges and a dual threshold obtained by different techniques. They obtain the control point for the snake with the comparison of the two regions. They evaluate the method over the MIAS database.

- ***Classifiers based techniques.*** Lou *et al* [22], used a clustering approach to obtain an initial region, estimates the real boundary extrapolating and linking those detected points. Saha *et al* [23] use a scale-based fuzzy connectedness algorithm. Rickard *et al* [24] presents Self-organizing map, a type of unsupervised artificial neural network model. The method applied by Wirth *et al* [25] was a fuzzy segmentation and evaluates the results in terms of completeness and correctness comparing the images from the MIAS database with a gold standard manually generated. Recently, Tromans *et al* [26], use a mixture model to obtain a mathematical representation of the image background and the compressed parameters, combined with a Fourier model, using an Expectation Maximization algorithm.

Summarizing, the traditional histogram based method has provided good and quick results. This quality sometimes turns on weakness in difficult cases where can be enhanced with local histogram or gradient approaches. The polynomial modelling and active contours provide very good results with accurate profiles.

2.1 Our Method

Figure 1 shows a visual scheme of the proposed method. To achieve the segmentation we propose a “two-phase” based method. It combines an adaptive Histogram approach to separate the breast from the background (Phase A), and a selective region growing algorithm to obtain pectoral muscle suppression (Phase B).

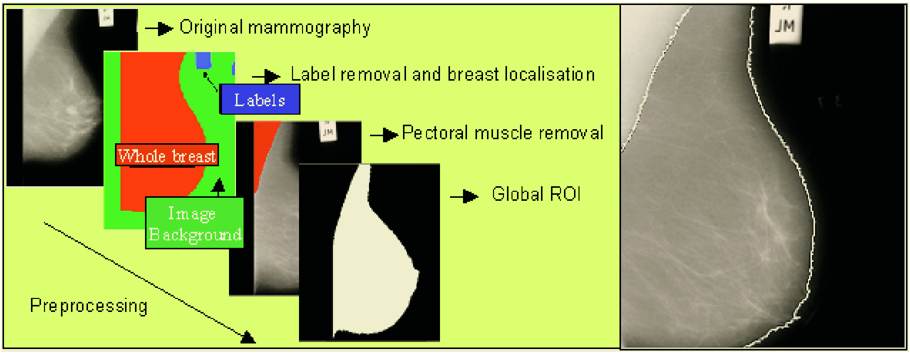


Fig. 1. Global Segmentation. Breast region extraction and pectoral muscle suppression.

Phase A. Breast Segmentation. Figure 1 shows the steps followed from the original mammogram to obtain the breast mask. A global histogram is calculated and smoothed with a gaussian operator. N consecutive percentage of bright pixels are tested to obtain N thresholds (ie. 10%, 15%, 20% that in grey level means 220, 210, 200). Each value is used in order to threshold the image and obtain masks which are overlapped. The region defined by the boundary of the smallest threshold to the boundary of the largest one is statistically evaluated to calculate the mean of the grey level which is used as our final threshold value. The result of applying this threshold is a collection of different regions. The largest one is the union of the breast and the pectoral muscle. We extract this largest region using a Connected Component Labelling algorithm [28]. In Figure 1, the region of interest of the breast has been extracted from the pectoral muscle using the region growing algorithm described above. In the following section we introduce a new method to detect the pectoral muscle using a selective region growing approach.

Phase B. Extracting the Pectoral Muscle. This operation is important in mediolateral oblique view (MLO), where the pectoral muscle, slightly brighter compared to the rest of the breast tissue, can appear in the mammogram.

Previous work related to pectoral muscle suppression used Hough Transform [14, 29], assuming that the boundary between the pectoral muscle and the breast can be approximated by a straight line. Other related works are the one of Yam *et al* [30] whose work introduces a curvature component to the Hough estimation and the work presented by Ferrari *et al* [31] who propose a polynomial modelling of the pectoral muscle. The method we propose is inspired in the proposal of Georgsson [32] and in summary it follows the three steps:

1. **Breast localisation and orientation.** To classify the mammogram as right or left breast, we compare both sides of the breast profile, and using the curvature detected in each one, it is straightforward to determine the orientation.
2. **Region growing intensity threshold estimation (RG).** Once the orientation is known, a seed is placed inside the pectoral muscle (the first pixel of the non-curved side). A statistical region growing algorithm (RG) grows from this seed to fill the whole region of the pectoral muscle. A size restriction has been applied to avoid a wrong growing. When the limit of growing is exceeded, the growing criteria is corrected. This correction is estimated from the histogram of the previous region grown, progressively decreasing the initial value of the growing criteria. Then the RG is restarted as shown in Figure 2. If a correct growing is not found in finite steps, the initial mask is provided as a result and the no existence of pectoral muscle is assumed.
3. **Boundary refinement.** Finally, the pectoral muscle is suppressed from the breast region, and a morphological operator is applied to refine the boundary.

3 Experimental Results

We have used the public database MiniMIAS [33] to test our method. It is a reduced version of the original MIAS Database (digitized at 50 micron pixel edge) that has been reduced to 200 micron pixel edge and clipped or padded so that every image is 1024x1024 pixels.

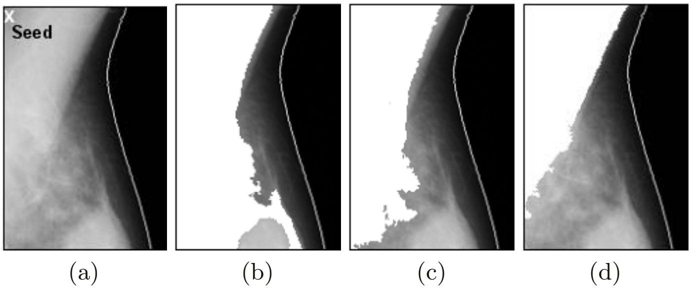


Fig. 2. Pectoral muscle removal. Region growing criteria correction. (a) original image,(b),(c) wrong RG (d) final correct RG.

Figure 3 shows three representative results. We have tested over 320 images, and we have obtained a 98% of "near accurate" results, which include the "accurate" results. About the muscle substraction, we have obtained a 86% of good extractions. Those results are obtained from a visual inspection of the images carried out by experienced radiologists and technicians trained with those kind of images. We should notice that some of them are a little bit over or under segmented. The behavior of the method shows an over-segmentation of the breast in cases with dense tissue, where the contrast between the muscle and the tissue is fuzzy. In that cases, our method rejects the muscle detection and provides the region obtained without suppressing the muscle as a final result. A possible solution could be to impose shape restrictions to the growing process. To summarize, the results obtained by the method show that it is a robust approach but it can be improved in terms of accuracy. Even so, we accept this method because it provides useful regions (there is no meaningful loss of information).

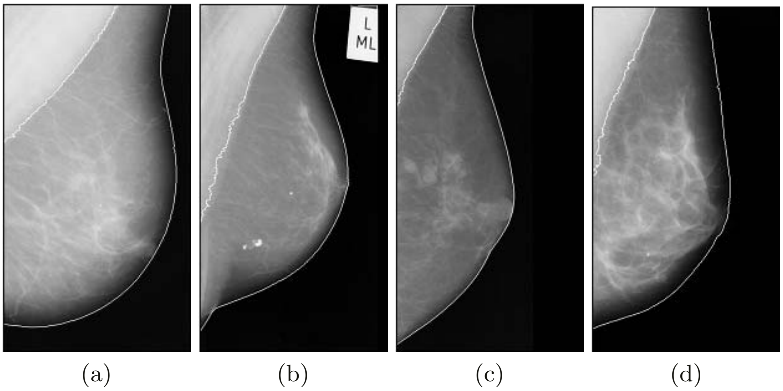


Fig. 3. An example of the performance of the presented approach on the segmentation of the profile of four different breasts.

4 Conclusions and Further Work

The literature survey will be a useful resource for others researching in this area. A new method to segment the breast with pectoral muscle suppression has been presented. The results obtained over MiniMIAS database have shown a general good behavior. In this sense we will focus our efforts to enhance the method as we consider that is important to take some shape features into account to deal with the more accurate pectoral muscle suppression. The results have shown that problems with the image acquisition, background noise, artifacts and scratches could all influence the reliability of the algorithm.

References

1. Reston, V., ed.: Breast Imaging Reporting and Data System. 4th edn. American College of Radiology (1998)
2. Hoyer, A., Spiesberg, W.: Computerized mammogram processing. In: Phillips Technical Review. Volume 38. (1979) 347–355
3. Lau, T., Bischoff, W.: Automated detection of breast tumors using the asymmetry approach. In: Computers and Biomedical Research. Volume 24. (1991) 273–295
4. Masek, M., Attikiouzel, Y.: Skin-air interface extraction from mammograms using an automatic local thresholding algorithm. In: ICB, Brno, CR (2000) 204–206
5. Yin, F., Giger, M.: Computerized detection of masses in digital mammogram: analysis of bilateral subtraction images. In: Medical Physics. Volume 28. (1991) 955–963
6. Bick, U., Giger, M.: Automated segmentation of digitized mammograms. In: Academic Radiology. Volume 2. (1995) 1–9
7. Byng, J., Boyd, N.: Automated analysis of mammographic densities. In: Medical Physics. Volume 41. (1996) 909–923
8. Hein, J., Kallargi, M.: Multiresolution wavelet approach for separating the breast region from the background in high resolution digital mammography. In: Digital Mammography, Nijmegen, Kluwer Academic Publishers (1998) 295–298
9. Semmlow, J., Shadagopappan, A.: A fully automated system for screening xeromammograms. In: Computers and Biomedical Reseach. Volume 13. (1980) 350–362
10. Méndez, A., Tahoces, P.: Automatic detection of breast border and nipple in digital mammograms. In: Computer Methods and Programs in Biomedicine. Volume 49. (1996) 253–262
11. Zhou, C., Chan, H.: Computerized image analysis: Estimation of breast density on mammograms. In: Med. Phys. Volume 28. (2001) 1056–1069
12. Abdel-Mottaleb, M., Carman, C.: Locating the boundary between the breast skin edge and the background in digitized mammograms. In: Digital Mammography. (1996) 467–470
13. Morton, A., Chan, H., Goodsitt, M.: Automated model-guided breast segmentation algorithm. In: Med Phys. (1996) 1107–1108
14. Karssemeijer, N., te Brake, G.: Combining single view features and asymmetry for detection of mass lesions. In: IWDM. (1998) 95–102
15. Stomatakis, E., Cairns, A.: A novel approach to aligning mammograms. In: Digital Mammography. (1994) 255–364

16. Chandrasekhar, R., Attikiouzel, Y.: Gross segmentation of mammograms using a polynomial model. In: International Conference of the IEEE in Medicine and Biology Society. Volume 3. (1996) 1056–1058
17. Goodsitt, M., Chan, H.: Classification of compressed breast shapes for the design of equalisation filters in x-ray mammography. In: Medical Physics. Volume 25. (1998) 937–947
18. Ojala, T., Liang, J.: Interactive segmentation of the breast region from digitized mammograms with united snakes. Technical Report 315, Turku Centre for Computer Science (1999)
19. Ferrari, R., Rangayyan, R.: Segmentation of mammograms: Identification of the skin boundary and the pectoral muscle. In: IWDM. Volume 23. (2000)
20. McLoughlin, K., Bones, P.: Locating the breast-air boundary for a digital mammogram image. In: Image and Vision Computing. (2000)
21. Wirth, M., Stapinski, A.: Segmentation of the breast region in mammograms using snakes. In: Canadian Conference on Computer and Robot Vision. (2004) 385–392
22. Gauch, J.: Image segmentation and analysis via multiscale watershed hierarchies. In: IEEE Transactions on Image Processing. Volume 8. (1999) 69–79
23. Saha, P., Udupa, J.: Breast tissue density quantification via digitized mammograms. In: IEEE Transactions on Medical Imaging. Volume 20. (2001) 792–803
24. Rickard, H., Tourassi, G., Elmaghraby, A.: Self-organizing maps for masking mammography images. In: IEEE EMBS. (2003) 302–305
25. Wirth, M., Lyon, J., Nikitenko, D.: A fuzzy approach to segmenting the breast region in mammograms. In: IEEE FI. Volume 1. (2004) 474–479
26. Tromans, C., Brady, J., Warren, R.: A high accuracy technique for breast air boundary segmentation and the resulting improvement from its use in breast density estimation. In: IWDM. (2004) 17–18
27. Ibrahim, N., Fujita, H.: Automated detection of clustered microcalcifications on mammograms: Cad system application to mias database. In: Physics in Medicine and Biology. Volume 42. (1997) 2577–2589
28. Sanz, L., Petkovic, D.: Machine vision algorithms for automated inspection of thin-film disk heads. Volume 10. (1988) 830–848
29. Kwok, S., Chandrasekhar, R., Attikiouzel, Y.: Automatic pectoral muscle segmentation on mammograms by straight line estimation and cliff detection. In: IIS Conference. (2001) 67–72
30. Yam, M., Brady, M.: Three-dimensional reconstruction of microcalcifications clusters from two mammographic views. In: Proc Medical Image. Volume 20. (2001) 479–489
31. Ferrari, R., Rangayyan, R.: Automatic identification of the pectoral muscle in mammograms. In: IEEE Transactions on Medical Imaging. Volume 23. (2004) 232–245
32. Georgsson, F.: Algorithms and Techniques for Computer Aided Mammographic Screening. PhD thesis, UMINF-01.15, Umeå University, Sweden (2001)
33. Suckling, J., Parker, J.: The mammographic images analysis society digital mammogram database. In: Exerpta Medica. International Congress Series. Volume 1069. (1994) 375–378

Semivariogram and SGLDM Methods Comparison for the Diagnosis of Solitary Lung Nodule

Aristófanes C. Silva[1], Anselmo C. Paiva[1],
Paulo C.P. Carvalho[2], and Marcelo Gattass[3]

[1] Federal University of Maranhão - UFMA,
Av. dos Portugueses, SN, Campus do Bacanga, Bacanga,
65085-580, São Luís, MA, Brazil
ari@dee.ufma.br, paiva@deinf.ufma.br

[2] Institute of Pure and Applied Mathematics - IMPA,
Estrada D. Castorina, 110, Horto, 22460-320,
Rio de Janeiro-RJ, Brazil
pcezar@impa.br

[3] Pontifical Catholic University of Rio de Janeiro - PUC-Rio,
R. Marquês de São Vicente, 225, Gávea,
22453-900, Rio de Janeiro, RJ, Brazil
mgattass@tecgraf.puc-rio.br

Abstract. The present work seeks to develop a computational tool to suggest the malignancy or benignity of Solitary Lung Nodules by means of analyzing texture measures obtained from computerized tomography images.Two methods are proposed, that analyze the nodules' texture by means of the Spatial Gray Level Dependence Method and a geostatistical function denominated semivariogram. A sample with 36 nodules, 29 benign and 7 malignant, was analyzed and the preliminary results of these methods are very promising in characterizing lung nodules. The obtained results suggested that the proposed methods have great potential in the discrimination and classification of Solitary Lung Nodules.

1 Introduction

Lung cancer is known as one of the cancers with shortest survival after diagnosis [1]. Therefore, the sooner it is detected the larger is the patient's chance of cure. On the other hand, the diagnosis accuracy grows with the amount of information that the physicians have available.

Solitary lung nodules are approximately round lesions less than 3 cm in diameter and completely surrounded by pulmonary parenchyma. Larger lesions should be referred to as pulmonary masses and should be treated under the assumption that they are most likely malignant - prompt diagnosis and resection are usually advisable [1].

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 479–486, 2005.

A number of techniques have been proposed to produce objective measures of Lung Nodules, based on texture, to ascertain malignancy or benignity. The most promising texture processing algorithms are mainly based on statistical parameters, such as the Spatial Gray Level Dependence Method-SGLDM, the Gray Level Difference Method-GLDM, and the Gray Level Run Length Matrices Method-GLRLM [2]. Other studies have involved the use of geostatistical parameters deduced from the variogram function as [3].

In this paper we analyze and compare the application of two methods based on texture (distribution attenuation coefficients) to the diagnosis of solitary lung nodules: the Spatial Gray Level Dependence Method and the Semivariogram. All the extracted measures of the methods are applied to the 3D nodules and are based on Computerized Tomography (CT) images.We consider nodules extracted from the original CT images using a 3D region-growing algorithm with voxel aggregation. The nodules' malignancy or benignity is determined by applying linear stepwise discriminant analysis and multilayer perceptron neural networks. The validation of the classifiers is done by means of the leave-one-out technique. The analysis and evaluation of tests are done using the area under the ROC curve.

This paper is organized as follows. Sections 2 and 3 discuss the studied method groups and classification techniques, respectively. Discussion and analysis of the results using the methods, stepwise discriminant analysis, multilayer perceptron neural networks and ROC curve are treated in Section 4. Finally, Section 5 presents some concluding remarks.

2 Texture Analysis Methods

The proposed methods analyzes nodules' texture by means of a classical image-processing method, the Spatial Gray Level Dependence Method (also called co-occurrence matrix), and by means of a geostatistical function called the semivariogram. The objective of these two methods is to obtain measures to discriminate benign lung nodules from malignant ones.

The Spatial Gray Level Dependence Method – SGLDM is a texture analysis technique that has been frequently used in 2D image segmentation and identification [4], [5] and [6]. Specific applications to medical images can be found in [2] and [5].

The SGLDM matrix is formed by tabulating the number of occurrences of each pixel gray level pair (i, j) that occurs within the ROI at a distance d along a direction defined by an angle θ. The choice of distance and angle combination, as well as the quantization level, is somewhat arbitrary.

To compute the SGLDM matrix in 3D, for each voxel we find all voxels gray level pairs that are at a distance d and acumulate their number of occurencies, then we normalize the matrix. As the SGLDM matrix dimension depends on the number of gray level pairs in the ROI, we perform a quantization in the ROI gray levels to reduce it and the associated noise.

For the nodule classification we used just 6 of the 13 measures proposed by Haralick et al. [6] to perform pattern recognition based on SGLDM matrix M.

The measures used in this work are: Contrast $(\sum_{i=0}^{G-1}\sum_{j=0}^{G-1} M_{i,j}(i-j)^2)$; Homogeneity $(\sum_{i=0}^{G-1}\sum_{j=0}^{G-1} \frac{M_{i,j}}{1+(i-j)^2})$; Angular second moment $(\sum_{i=0}^{G-1}\sum_{j=0}^{G-1} M_{i,j}^2$; Entropy $(-\sum_{i=0}^{G-1}\sum_{j=0}^{G-1} M_{i,j}\log(M_{i,j}))$; Variance: $\sum_{i=0}^{G-1}\sum_{j=0}^{G-1} (i-\mu)^{2} M_{i,j}$, where μ represents M's mean; and Correlation $(-\sum_{i=0}^{G-1}\sum_{j=0}^{G-1} M_{i,j}\left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}}\right]$, where μ is the mean and σ is the standard deviation of M.

Semivariance is a measure of the degree of spatial dependence between samples, depending on the distance between them. The plot of the semivariances as a function of distance from a point is referred to as semivariogram. The semivariogram summarizes the strength of associations between responses as a function of distance, and possibly direction [7]. Typically we assume that spatial autocorrelation does not depend on where a pair of observations (in our case, voxel or attenuation coefficient) is located, but only on the distance between the two observations, and possibly on the direction of their relative displacement.

The semivariogram is depicted by its sill, range and nugget. Sill is the asymptotic value; range is the distance at which this asymptotic value occurs; and nugget is the semivariance at a distance 0.0, that is, the intercept. It is defined by

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} (x_i - y_i)^2 \tag{1}$$

where h is the lag (vector) distance between the head value (source voxel), y_i, and the tail value (target voxel), x_i, and $N(h)$ is the number of pairs at lag h.

When computing directional semivariograms in 3D we define the azimuth(rotation in the horizontal plane) and dip (rotation from the horizontal plane) angles as the direction vector.

3 Classification Algorithms

A wide variety of approaches has been taken towards the classification task. This section provides an overview of Fisher's Linear Discriminant Analysis (FLDA) and Multilayer Perceptron based on the statistical and neural network paradigms.

The FLDA approach looks for linear combinations of the input variables that can provide an adequate separation for the given classes. The discriminant functions used by LDA are built up as a linear combination of the variables that seek to somehow maximize the differences between the classes:

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \beta^{'} x \tag{2}$$

The problem then is reduced to finding a suitable β vector. The method searches for a linear function in the attribute space that maximizes the ratio of the between-group sum-of-squares (B) to the within-group sum-of-squares (W). This can be achieved by maximizing the ratio $\frac{\beta' B \beta}{\beta' W \beta}$ and it turns out that the vector that maximizes this ratio, β, is the eigenvector corresponding to the largest eigenvalue of $W^{-1}B$. Hence the discriminant rule can be written as:

$$x \in i \;\text{ if }\; \left|\beta^T x - \beta^T u_i\right| < \left|\beta^T x - \beta^T u_j\right|, \text{for all } \; j \neq i \tag{3}$$

where $W = \sum n_i S_i$ and $B = \sum n_i (x_i - x)(x_i - x)^{'}$, and n_i is class i sample size, S_i is class i covariance matrix, x_i is the class i mean sample value and x is the population mean.

Stepwise discriminant analysis was used to select the best variables to differentiate between groups. These measures were used by Fisher's Linear Discriminant Analysis and by Multilayer Perceptron Neural Networks classifiers.

The Multilayer Perceptron - MLP, a feed-forward back-propagation network, is the most frequently used neural network technique in pattern recognition [8]. Briefly, MLPs are supervised learning classifiers that consist of an input layer, an output layer, and one or more hidden layers that extract useful information during learning and assign modifiable weighting coefficients to components of the input layers. During training, MLPs construct a multidimensional space, defined by the activation of the hidden nodes, so that the two classes (benign and malignant nodules) are as separable as possible. The separating surface adapts to the data.

In order to validate the classificatory power of the discriminant function, the leave-one-out technique [9] was employed. Through this technique, the candidate nodules from 35 cases in our database were used to train the classifier. The trained classifier was then applied to the candidate nodules in the remaining case. This technique was repeated until all 36 cases in our database had been the "remaining" case.

In order to evaluate the ability of the classifier to differentiate benign from malignant nodules, we used the area (AUC) under the ROC curve (Receiver Operating Characteristic) [10].

4 Results

This section shows the results of applying the two proposed methods of texture characteristics extraction for a set of nodules and the results obtained from the use of discriminant analysis and multilayer perceptron classifiers to the classification of the nodules between malignant and benign classes.

In the test we used images acquired from an Helical GE Pro Speed tomography, with resolution of 512×512 pixels, voxel size $0.67 \times 0.67 \times 1.0$ mm. The images were quantized in 12 bits and stored in the DICOM format.

The tests described in this paper were carried out using a sample of 36 nodules, 29 benign and 7 malignant. It is important to note that the nodules were diagnosed by physicians and had the diagnosis confirmed by means of surgery or based on their evolution. Such process takes about two years, which explains the reduced size of our sample. The sample included nodules with varied sizes and shapes, with homogeneous and heterogeneous characteristics, and in initial and advanced stages of development.

Figure 1 show the application of the SGLDM method, we can notice that benign nodules have a more distributed voxel density than the malignant one.

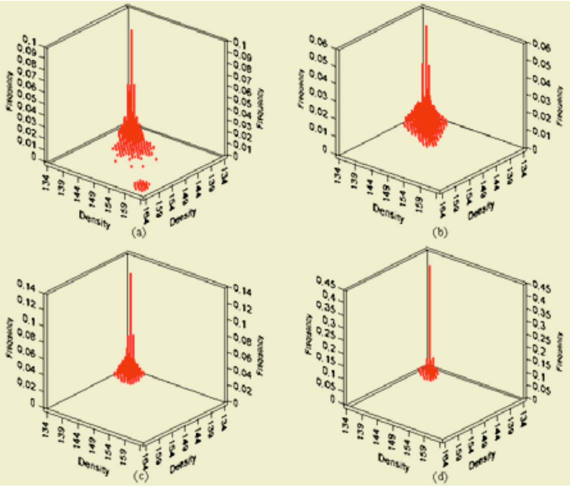


Fig. 1. SGLDM method applied to the test nodules.

Figure 2 shows the application of experimental semivariograms to test nodules. We can notice that benign nodules have a higher sill than malignant ones, and that their initial slope is much more accentuated. The graph analysis shows the presence of more dispersion in benign nodules than in malignant ones.

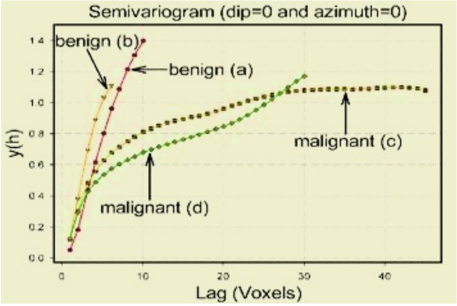


Fig. 2. Semivariogram applied to the test nodules.

Figures 3 and 4 exemplify, respectively, the application of the semivariogram function to a benign nodule and to a malignant nodule. The curves show the computed variance in twelve specified directions (dip equal to 0°,−45° and −90°, and azimuth equal to 0°, 45°, 90°, and 135°), in relation to several distances. We can conclude from the results that the semivariogram function presents isotropic characteristics in the tests undertaken.

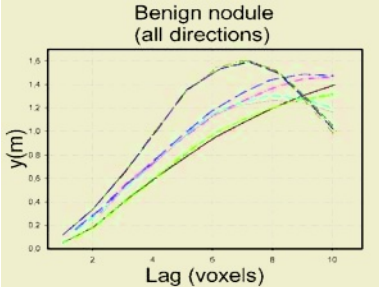


Fig. 3. Semivariogram applied to the test nodules.

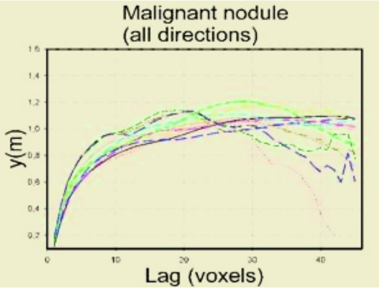


Fig. 4. Semivariogram applied to the test nodules.

We used 26 neighbors in the SGLDM and tested the method with a distance of 1, 2 and 3 voxels and number of gray levels of 8, 16, 32, 64 and 256. Thus, we have 90 measures for the SGLDM method.

The stepwise discriminant analysis selected 9 out of 90 measures to be analyzed by the FLDA and MLP classifiers. By analyzing these selected measures, some considerations can be made: i) only one of the selected measures was quantized with 8 gray levels; this is explained by the fact that the smaller the number of gray levels, the larger the amount of information lost; ii) only one selected measure was quantized with 256 gray levels, because the larger the number of gray levels, the sparser is the resultant matrix or histogram, and consequently the analysis becomes statistically poor.

Table 1 shows the results of SGLDM method based on the studied classifiers. The areas under the ROC curve for the two classifiers were above 0.9, which means results with excellent accuracy. There is not a statistically significant difference between the ROC curves of the two classifiers (p-value = 0.641).

In the semivariogram study, analytical models for the geostatistical function were not used; instead, empirical geostatistical function was employed. The measures (variables) extracted, considered as texture signatures, were obtained by computing the semivariogram function for a set of directions: dip (Z) 0°,−45°, and −90°. For each dip the azimuth (X and Y) is 0°, 45°, 90°, and 135°. The adopted lag separation distance (h) was 1 mm, with tolerance angle of ±22.5°, and tolerance lag of ±0.50 mm. The maximum number of lags depends on the size of each image (volume). We have selected the first three and the last lags (h) in a specific direction for each function. They were selected because we were interested in verifying slight variations in small distances, but without rejecting the information of larger distances. In total, we have 36 measures (3 dips × 4 azimuths × 4 lags) for the semivariogram.

Table 1. Accuracy analysis in the diagnosis of lung nodules.

Method	Classifier	Benign (%)	Malignant (%)	Accuracy (%)	AUC
SGLDM	FLDA	89.7	71.4	86.1	0.946
	MLP	89.7	85.7	88.8	0.906
Semivariogram	FLDA	93.1	100.0	94.4	1.0
	MLP	96.5	100.0	97.2	1.0

The stepwise technique selected 5 from the 36 measures to be analyzed by the FLDA and MLP classifiers. By analyzing these selected measures, some considerations can be made: i) no selected measure had lag equal to 1, which contradicts our initial idea that using small lags we will provide details of the lung nodules; ii) four selected measures had dip with value different from 0°, showing that the nodules' 3D characteristics are essential to their discrimination and classification.

Table 1 shows the results of semivariogram function based on the studied classifiers. Analyzing the area under the ROC curve, we have observed that the two classifiers have $AUC = 1.0$, which means results with excellent accuracy. Of course, there is not a statistically significant difference between the ROC curves of the two classifiers (p-value = 1.0).

The results show that the analysis of individual and combined groups provides an accuracy of over 80% in the diagnosis of lung nodules. Moreover, we can observe the following: i) there isn't a great predominance of one of the two classifiers; ii) all tests performed demonstrated that the accuracy in the diagnosis is considered excellent; iii) semivariogram function have $AUC = 1.0$, considered perfect, which it isn't present in the SGLDM.

Due to the relatively small size of the existing CT lung nodule databases and the various CT imaging acquisition protocols, it is difficult to compare the diagnosis performance between the developed algorithms and others proposed in the literature.

5 Conclusion

This paper has presented two methods to analyze the nodules' texture by means of the Spatial Gray Level Dependence Method and semivariogram function with the purpose of characterizing lung nodules as malignant or benign. The measures extracted from SGLDM method and semivariogram function were analyzed and had great discriminatory power, using discriminant analysis to classify and the ROC curve to evaluate the obtained results. The number of nodules studied in our dataset is too small and the disproportion in the samples does not allow us to make definitive conclusions, However, the results obtained with our sample are very encouraging, demonstrating that the linear discriminant and the multilayer perceptron classifiers using characteristics of the nodules' texture can effectively classify benign from malignant lung nodules based on CT images. Nevertheless, there is the need to perform tests with a larger database and more complex cases in order to obtain a more precise behavior pattern.

Despite the good results obtained only by analyzing the texture, further information can be obtained by analyzing the geometry. As a future work, we propose a combination of texture and geometry measures for a more precise and reliable diagnosis.

Acknowledgments

We thank Dr. Rodolfo Nunes and his team for the clinical support, and Dr. Marcia Boechat with the staff of Instituto Fernandes Figueira, for the images.

References

1. Tarantino, A.B.: 38. In: Nódulo Solitário Do Pulmão. 4 edn. Guanabara Koogan, Rio de Janeiro (1997) 733–753
2. McNitt-Gray, M.F., Hart, E.M., Wyckoff, N., Sayre, J.W., Goldin, J.G., Aberle, D.R.: A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: Preliminary results. Medical Physics **26** (1999) 880–888
3. Bruniquel-Pinel, V., Gastellu-Etchegorry, J.P.: Sensitivity of texture of high resolution images of forest to biophysical and acquisition parameters. Remote Sensing of Environment **65** (1998) 61–85
4. Jain, A.K.: Fundamentals of Digital Image Processing. Prentice Hall, Englewood Cliffs, NJ, USA (1989)
5. McNitt-Gray, M.F., Hart, E.M., Wyckoff, N., Sayre, J.W., Goldin, J.G., Aberle, D.R.: The effects of co-occurrence matrix based texture parameters on the classification of solitary pulmonary nodules imaged on computed tomography. Computerized Medical Imaging and Graphics **23** (1999) 339–348
6. Haralick, R., Shanmugam, K., Dinstein, I.: Textural features for image classification. SMC **3** (1973) 610–621
7. Journel, A.G., Huijbregts, C.J.: Mining Geostatistics. Academic Press, London (1978)
8. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, New York (1999)
9. Fukunaga, K.: Introduction to Statistical Pattern Recognition. 2 edn. Academic Press, London (1990)
10. Erkel, A.R.V., Pattynama, P.M.T.: Receiver operating characteristic (ROC) analysis: Basic principles and applicattions in radiology. European Journal of Radiology **27** (1998) 88–94

Anisotropic 3D Reconstruction and Restoration for Rotation-Scanning 4D Echocardiographic Images Based on MAP-MRF

Qiang Guo[1], Xin Yang[1], Ming Zhu[2], and Kun Sun[2]

[1] Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University
Shanghai 200030, P.R. China
{guoqiang1221,yangxin}@sjtu.edu.cn
[2] Xinhua Hospital, Attached to Shanghai Second Medical University
Shanghai 200092, P.R. China
zhuming58@vip.sina.com, sunkun@hotmail.com

Abstract. An anisotropic method of 3D reconstruction method for time sequence echocardiographic images is proposed in this paper. First, a Bayesian model based on MAP-MRF is described to reconstruct 3D volume, and extended to deal with the images acquired by rotation scanning method. Second, the spatial and temporal nature of ultrasound images is taken into account for the selection of parameter of energy function, which makes this statistical model anisotropic. Hence not only can this method reconstruct 3D ultrasound images, but also remove the speckle noise anisotropically. Finally, we illustrate the experiments of our method on the synthetic and medical images and compare with the isotropic reconstruction method.

1 Introduction

Three-dimensional echocardiography has been widely used in diagnostic cardiology because it can visualize the complex structure of heart more accurately than previous used 2D diagnosis method. Currently, there are two major images acquisition methods of 3D echocardiography, i.e. random data acquisition and sequential data acquisition [1]. The former, which allows unrestricted (free-hand) scanning from any available precordial acoustic window, uses a spatial locator to measure the position and orientation of the ultrasound probe. It is usually adopted to assess the structure and surface shapes and to analyze the left ventricular volume quantitatively. In our work, three-dimensional echocardiography is applied to analyze the abnormality of the mitral valve, which causes complex congenital heart disease. In this application, rotational scanning, one mode of sequential data acquisition, is used to acquire the ultrasound images. This approach is shown in Fig.1. The end-firing endocavity probe is put on the skin near heart (see Fig.1 (a)). Each acquisition consists of 13-17 frames per cardiac cycle depending on the heart rate. The angular slice spacing was 3 degrees resulting in 60 images slices in the acquisition process. Therefore, the acquired images are time series of 3D echocardiographic images, i.e. 4D images. Each 3D ultrasound images represents a state of heart at one time instant during the cardiac cycle (see Fig.1 (b)). The acquired images intersect each other along a revolution axis forming a cylindrical geometry, which may be used to reconstruct the 3D structure of heart (see Fig.1(c)).

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 487–494, 2005.

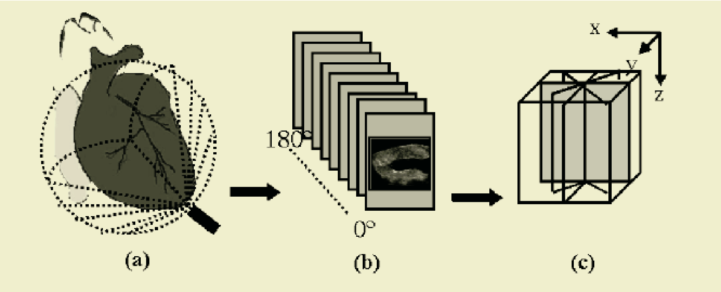


Fig. 1. Rotation-scanning approach with end-firing probe

There are two distinct reconstruction approaches [2]: feature-based and voxel-based reconstruction. In the feature-based approach, the desired features (e.g. contours or surfaces) of anatomical structures are first determined and then reconstructed into a 3D image. The second approach is based on using a set of acquired 2D images to build a voxel-based image, i.e. a regular Cartesian grid of volume elements in three dimensions. For each 3D image point, the voxel value (intensity) is calculated by interpolation, as the weighted average of the pixel values of its nearest neighbors among the embedded 2D image pixels. In 1982, Ghosh carried out the preliminary study on 3D reconstruction of cylindrical echocardiographic images [3]. Following his work, many different interpolation methods were proposed. For example, Nearest-neighbor interpolation, linear interpolation and bilinear interpolation were used in [4]. The concentric interpolation was introduced by Duann [5]. Moreover, the statistical method formulated in Bayesian framework may also be applied in the interpolation [6].

Sanches introduced a Bayesian approach, which uses the Rayleigh distribution as the observation model, to reconstruct the 3D ultrasound images [7]. Not only can this approach reconstruct the 3D volume, but also remove the speckle noise simultaneously. Based on the same framework, he proposed a statistical approach [8] to compensate the misalignment and geometric distortions of the acquisition images. Furthermore, Sanches proposed a fast algorithm to speed up the procedure of solving those problems [9]. However, his work is based on the images acquired by free-hand method, which is quite different from the rotation scanning method geometrically. Inspired by his work, we present an anisotropic 3D reconstruction and restoration algorithm designed for rotation-scanning ultrasound images. First, a MAP-MRF based reconstruction method is established to reconstruct 3D volume, using the images acquired by rotational scanning method. Then the spatial and temporal nature of ultrasound images is introduced for the selection of parameter α in the energy function. Parameter α is the measurement of the strength of the connection between neighborhood nodes. Hence this reconstruction model is anisotropic, which avoid blurring the boundary during the reconstruction process.

2 Reconstruction Model

As proposed by Sanches [7], in the method of voxel-based 3D reconstruction, the 3D space is divided into cubic cells. The value of a voxel is the weighted average of the pixel values of its nearest neighbors, i.e. u_1, u_2…u_8 (see Fig.2 (a)). Hence the intensity of a voxel may be considered to be the value of function f(x), i.e.

$$f(x) = B(x)^T U \tag{1}$$

where B(x)={b_i(x)} is a vector of basis functions and U={u_i}is a vector of unknown coefficients to be estimated. It is assumed that each b_i(x) is a basis function obtained by shifting a tri-linear interpolation function. Calculating u_i requires a finite support region (See Fig.2 (b)). u_{i1},u_{i2},u_{i3},u_{i4} are on one piece of 2D ultrasound image and u_{i5},u_{i6},u_{i7},u_{i8} are on the other. There is angular slice spacing between the two adjacent ultrasound images. These eight nodes compose the finite support region. How to estimate U from those known nodes is the key step in the reconstruction.

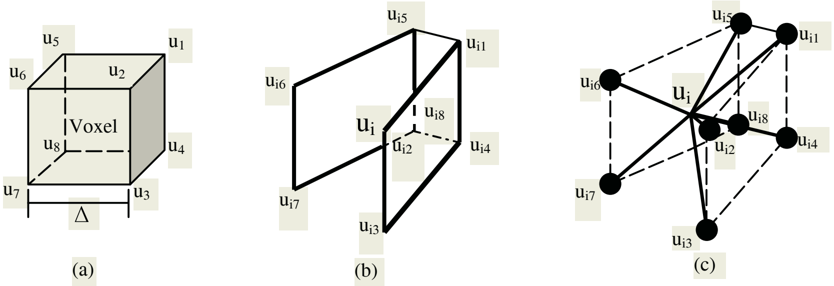


Fig. 2. (a) shows one voxel in the cubic regular grid. Each neighbor of the voxel can be calculated according to the structure shown in **(b), (c)** shows the neighborhood representation

The MAP-MRF framework [10] is applied to estimate the coefficient vector U, i.e.

$$\hat{U} = \arg\max \ln(p(Y \mid U)p(U)) \tag{2}$$

where Y={y_i, x_i}is the available data, y_i is the intensity of location x_i, p(Y|U)is the likelihood function and p(U) is the prior. Formulation of these two terms is as follows: Pixel value (i.e., intensity) can be considered as realization of independent random variables with Rayleigh distribution. So the likelihood function may be written as:

$$p(Y \mid U) = \prod_i \frac{y_i}{f(x_i)} e^{-(y_i^2/2f(x_i))} \tag{3}$$

where f(x_i) is the value of the function f(x) to be reconstructed at x_i .

The interpolation information is included in the prior p(U). According to the spatial properties of human organs and tissues, a Gibbs prior is used [11]. The prior distribution is given by:

$$p(U) = Z^{-1} \times e^{-\alpha \sum_{p \in G} \sum_{i \in N(p)} (u_p - u_i)^2} \tag{4}$$

where Z is a normalization factor and G denotes the gird nodes and N(p) are the neighbors of the p-th node on the given 2D images. The adoption of this prior is equivalent to consider that the neighboring should have similar values.

The neighborhood is defined on the irregular sites, i.e. the wedged grid (see Fig.2(c)). Each grid node is connected to eight neighbors, except boundary nodes. The parameter α measures the strength of each connection; It plays a very important role in the algorithm of reconstruction. Discussion about the parameter α is shown in the section 3.

MAP criteria require the maximization of the joint density L(U) with respect to U.

$$L(U) = \ln p(Y,U) = \ln p(Y \mid U)p(U)$$
$$= \ln(\prod_i \left[\frac{y_i}{f(x_i)} e^{y_i^2/2f(x_i)} \right] \frac{1}{Z} e^{-\alpha \sum_{(i,j)} (u_i - u_j)^2}) \tag{5}$$

Using Iterated Conditional Model (ICM) algorithm [12] to solve the problem of the maximization of L(U) with respect to U, which leads to

$$u_t = \frac{1}{4\alpha N} \sum_i \frac{y_i - 2f(x_i)}{f^2(x_i)} b_p(x_i) + \overline{u}_t \tag{6}$$

where u_t is the intensity of a grid node, N is the number of neighbors of u_t and $\overline{u}_t$ denotes the average intensity of the neighbors of u_t. Eq.(6) suggests an iterative procedure to compute u_t.

3 Parameter Formulation

The adoption of Gibbs distribution in the prior introduces the regularization effect while searching for a global minimum. However it convert the objective function to a non-convex one. So, general method cannot guarantee to obtain the global solution. Therefore, the careful selection of statistical parameter in prior plays a very important role. The bigger the value of parameter α is, the lower the convergence rate of the algorithm is, because an increased dependence is enforced among neighboring nodes. Therefore the parameter α can't remain a constant during the optimization process, or the algorithm will be trapped in a local minimum, far from the desired solution. Sanches demonstrated how the parameter α changes during the 3D reconstruction process [13].

On the other hand, the parameter α reflects the intensity difference between the neighboring nodes. High value of α is corresponding to a strong connection between two neighboring nodes, while low values of α corresponding to weak connections. In this paper, we take into account the connection of the nodes for the selection of parameter α , in addition to its influence on the convergence of the algorithm.

The strength of connection between the neighboring nodes can be measured by spatial gradient. A large spatial gradient means the weak connection, which often occur in the area of boundary; while the small value of gradient represents the strong connection of neighboring nodes, which indicates the local smoothness. Furthermore, the acquired images are time sets of data. So the temporal consistency can be also introduced to the selection of parameter α . The formulation of the selection of parameter α is shown as follows.

$$\alpha = \alpha_{ini} \times \beta \tag{7}$$

where α_{ini} is the initial value resulting in the isotropic volume; β is the adaptative factor, whose value is the result of penalty function S(x). It may be defined by:

$$\beta = S(x) = \begin{cases} arcctg(x-\eta)+\pi, & x<\eta \\ arcctg(x-\eta), & x \ge \eta \end{cases} \tag{8}$$

where η is a large positive constant. The plot of penalty function is shown in Fig.3. The input value of penalty function is given by the gradient measurement, which is represented by the x-axis in the Fig.3; the y-axis of function plot represents the adaptative factor, which is the output value of the penalty function. The function curve in Fig.3 indicates that the value of adaptative factor is inverse proportion to the value of the gradient measurement. The gradient measurement is given by g(x,y,z,t):

$$g(x,y,z,t) = E \cdot \left[D\nabla I(x,y,z,t) \right] \tag{9}$$

where E={1,1,1,1}, I(x,y,z,t) is the 3D ultrasound image at instant t. D denotes the anisotropic tensor. It can be written as:

$$D = diag(k(\frac{\partial I}{\partial x},\delta_x), k(\frac{\partial I}{\partial y},\delta_y), k(\frac{\partial I}{\partial z},\delta_z), k(\frac{\partial I}{\partial t},\delta_t)) \tag{10}$$

In the above equation, $k(x,\delta)$ is the diffusion function [14]. parameter δ denotes the gradient threshold. This function has the effect of reducing the diffusion at ‘high’ gradients, based on a threshold δ .

From (9) and (10), we can deduce the following expression:

$$\begin{aligned} g(x,y,z,t) = k(\frac{\partial I}{\partial x},\delta_x)\frac{\partial I}{\partial x} + k(\frac{\partial I}{\partial y},\delta_y)\frac{\partial I}{\partial y} \\ + k(\frac{\partial I}{\partial z},\delta_z)\frac{\partial I}{\partial z} + k(\frac{\partial I}{\partial t},\delta_t)\frac{\partial I}{\partial t} \end{aligned} \tag{11}$$

The expression above is discretized by the finite differences method. Since the given nodes is on the irregular grid, we approximate the partial differential as follows:

$$\frac{\partial I}{\partial x} = \left\| \frac{1}{4\Delta_x} \left[(u_{i1}+u_{i2}+u_{i5}+u_{i6}) - (u_{i3}+u_{i4}+u_{i7}+u_{i8}) \right] \right\| \tag{12}$$

$$\frac{\partial I}{\partial y} = \left\| \frac{1}{4\Delta_y} \left[(u_{i1}+u_{i2}+u_{i3}+u_{i4}) - (u_{i5}+u_{i6}+u_{i7}+u_{i8}) \right] \right\| \tag{13}$$

$$\frac{\partial I}{\partial z} = \left\| \frac{1}{4\Delta_z} \left[(u_{i2}+u_{i3}+u_{i6}+u_{i7}) - (u_{i1}+u_{i4}+u_{i5}+u_{i8}) \right] \right\| \tag{14}$$

$$\frac{\partial I}{\partial t}=\left\|\frac{1}{8\Delta_t}\left(\sum_{j=1}^{8}u_{ij}^{t+1}-\sum_{j=1}^{8}u_{ij}^{t}\right)\right\| \tag{15}$$

where $\mathrm{u}_{\mathrm{ij}}^{\mathrm{t}}$ denotes the intensity of node u_{ij} at instant t; Δ_{x} , Δ_{y} and Δ_{z} are grid step in three directions. Δ_t is the time step.

For each wedged gird, Eq.(6) is implemented to calculate the coefficient u_i. Before this iterative procedure, parameter α may be gained by the approach present above. Then, it reduced gradually during the iterative process as shown in [13].

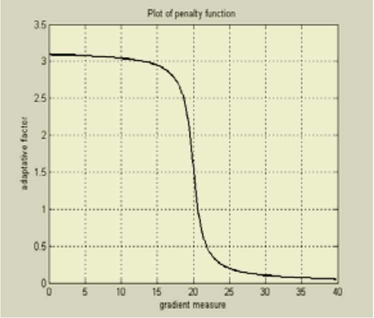


Fig. 3. Plot of penalty function S(x)

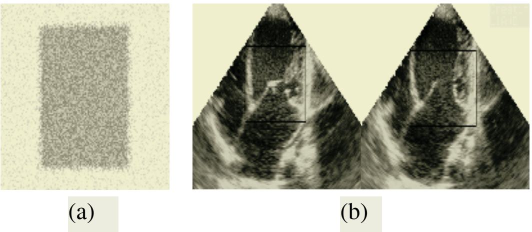

(a)
(b)

Fig. 4. Source images, **(a)** shows the synthetic slice and **(b)** shows the real ultrasound images

4 Experimental Results

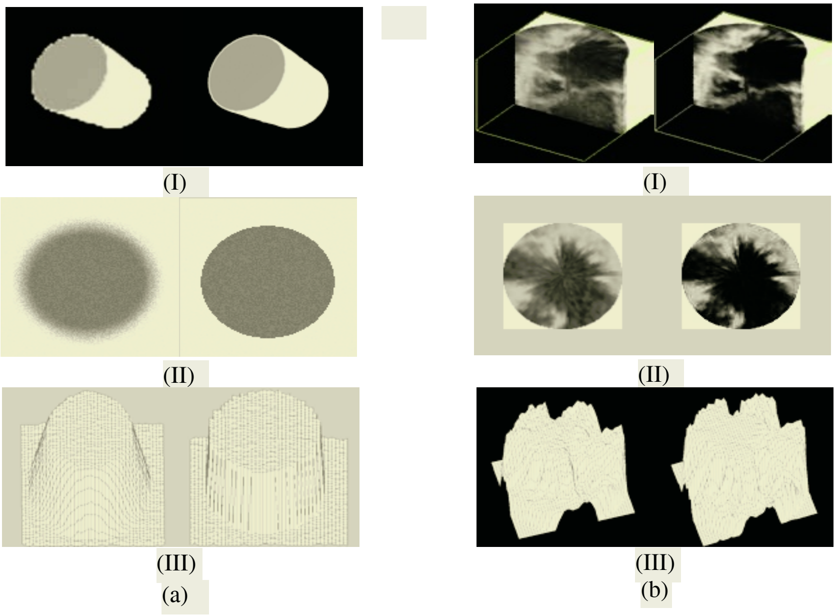
The anisotropic reconstruction algorithm introduced in the previous sections was tested on the synthetic and medical images, comparing with the isotropic reconstruction method.

The synthetic data consists of a set of 60 images of 128*128 pixels, each of them intersect each other along a revolution axis and the angular slice spacing of adjacent images is 3 degrees. One of those images is shown in Fig.4(a). It is blurred by Gaussian low pass filter and contaminated by the speckle noise. So, all those synthetic images simulate the real ultrasound images acquired by rotational scanning method.

A grid of 127*127*127 nodes was used in the test. Experimental results are shown in Fig.5(a), left column of those figures are the result of isotropic reconstruction method, right column of those figures are from anisotropic approach. Both two 3D reconstructed volumes are shown in Fig.5(a)(I). In order to make the difference between the two results more visible, two cross-sections are extracted from the estimated volume respectively, as shown in Fig.5(a)(II). Moreover Fig.5(a)(III) shows the surface plots of these two cross-sections. Those figures prove that our reconstruction method has edge-preserving ability.

Similar experiments were performed with the medical images, which were acquired by Philips 5500 TTO probe. The first two frames of source images is shown as

Fig.4(b). Those two figures are adjacent images with 3 degrees angular spacing. Since reconstructing the whole images consume much time, we only reconstructed the region of interest with black rectangle as shown in Fig.4(b), i.e. the mitral valve.



(I) (I)

(II) (II)

(III) (III)

(a) (b)

Fig. 5. Experimental results

A grid of 84*84*84 nodes was chosen to reconstruct the volume. A comparison of estimated volume between two algorithms is shown in Fig.5(b). Left column of those figures are the result of isotropic reconstruction method; right column of those figures are from anisotropic approach. The two reconstructed volumes are shown in Fig.5(b)(I). 3D reconstruction is the precondition for the segmentation of the mitral valve. Fig.5(b)(II) shows two cross-sections extracted from the estimated volume. The surface plots of those two cross-sections are shown in Fig.5(b)(III), which indicate the difference between the two methods.

5 Conclusion

3D reconstruction and restoration is a very important step in the quantitative analysis of mal-function of mitral valve. In this paper, we propose an anisotropic algorithm based on MAP-MRF for this purpose. There are three advantages of this method. First, the anisotropic property, which results in a clear boundary in the estimated volume, benefits the succeeding segmentation step in our work. Second, the statistical method is easily extended to deal with the artifacts in the echocardiograpic images. Third, during the process of 3D reconstruction, the speckle noise can be removed effectively due to the adoption of Rayleigh distribution in the likelihood function.

Acknowledgment

This work was partially supported by National Basic Research Program of China (No.2003CB716104) and Shanghai Science and Technology Research Foundation (No.034119820).

References

1. R.T.C.Jos, Roelandt, Three-dimensional echocardiography: the future today!, Computers & Graphics, vol.24, pp.715-729, 2000.
2. Aaron Fenster, Donal B., Three-dimensional ultrasound imaging Phys, Med.Biol., vol.46, pp.67-99, 2001.
3. Amitabha Ghosh, Navin C., Nanda, Gerald Maurer, Three-dimensional reconstruction of echocardiographic images using the rotation method, Ultrasound in Med.&Biol., vol.8,no.6, pp.655-657.
4. Tong S., D.B.Downey, H.N.Cardinal, A.Fenster, A Three-dimensional ultrasound prostate imaging system, Ultrasound Med.&Biol., vol.22, no.6, pp.735-746, 1996.
5. J.R.Duann, S.B.Lin et al, Computer system for four-dimensional transesophageal echocardiographic image reconstruction, Computerized Medical Imaging and Graphics, vol.23, pp. 173-179,1999.
6. Chun Jen Tsai, Yi Ping Hung, Shun Chin Hsu, Comparison between Asymptotic Bayesian Approach and Kalman Filter-Based Technique for 3D Reconstruction Using an Image Sequence, CVPR '93. IEEE Computer Society Conference on, pp.206 – 211.
7. J.M.Sanches, J.S.Marques, A Rayleigh reconstruction/interpolation algorithm for 3D ultrasound, Pattern Recognition Letter, vol.21, pp. 917-926,2000.
8. J.M.Sanches, J.S.Marques, Joint image registration and volume reconstruction for 3D ultrasound, Pattern Recognition Letter, vol.24, pp. 791-800,2003.
9. J.M.Sanches, J.S.Marques, A fast MAP Algorithm for 3D Ultrasound, EMMCVPR 2001, pp.63-74.
10. German, S., German, D., Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images, IEEE Trans. on PAMI., vol.6, no.6, pp.721-741,1984.
11. S.Z.Li, Markov Random Field Modeling in Computer Vision 2nd, Springer-Verlag, pp.37-63, 2001.
12. Besag J., On the statistical analysis of dirty pictures, Journal of the Royal Statistical Society B, vol.48, no.3, pp. 259--302, 1986.
13. J.M.Sanches, J.S.Marques, A Multiscale Algorithm For Three-Dimensional Free-hand Ultrasound, Ultrasound in Med.&Biol., vol. 28, no.8, pp.1029-1040, 2002.
14. Weickert J., Anisotropic diffusion in image processing, Teubner Verlag, 1998.

Comparison of Feature Extraction Methods for Breast Cancer Detection

Rafael Llobet, Roberto Paredes, and Juan C. Pérez-Cortés

Instituto Tecnológico de Informática*
Universidad Politécnica de Valencia
Camino de Vera, s/n 46071 Valencia, Spain
{rllobet,rparedes,jcperez}@iti.upv.es

Abstract. Although screening mammography is widely used for the detection of breast tumors, it is difficult for a radiologist to interpret correctly a mammogram. It is possible to improve this task by using a computer aided diagnosis system (CAD) which highlights the areas most likely to contain cancer cells. In this paper, we present and compare five different feature extraction methods for breast cancer detection in digitized mammograms. All the methods are based on features extracted from a local window and on a k-nearest neighbor classifier with fast search.

1 Introduction

Mammographic screening programs are currently an effective method to detect breast cancer at an early stage. Nevertheless, it is difficult for a radiologist to interpret correctly a mammogram due to the extremely wide variation in the mammographic appearance of normal and abnormal tissue of the breast. Retrospective studies have shown that, in current breast cancer screening, between 10% and 25% of the tumors are missed by the radiologist [19, 20]. While false positive errors may result in an unnecessary biopsy, false negative errors may result in delayed diagnosis of an actual cancer.

There is a clear need for methods of automatic detection of suspicious tissue in mammograms. Currently, automated analysis of digital mammograms is an active research field which aims to help radiologists to improve the diagnosis, reducing both the number of tumors that are missed (false negatives) and the number of biopsies of healthy tissue (false positives).

Following this idea, different schemes of feature extraction and classification methods have already been used to detect and classify regions of interest (ROI) in medical images in general and breast tumors in digitized mammograms in particular. Among the feature extraction approaches used in the literature, Laplacian of Gaussian filtering, template matching, methods based on gradient [3], space gray level dependence matrices (SGLDM) [2, 6, 16], independent

* This work has been partially supported by the Spanish CICYT under grant TIC2003-08496-CO4 and by *Agencia Valenciana de Ciencia y Tecnología* under contract GRUPOS03/031.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 495–502, 2005.

component analysis (ICA) [6], fractal analysis [7], support vector machines [4] and wavelets [13] have been commonly used in tissue segmentation. On the other hand, numerous classification methods have been tested, which are based on Markov models [16], neural networks [6] and k-nearest neighbors (k-NN) [16].

In this paper, a comparison of different feature extraction methods for automatic detection of tumors in digitized mammograms is presented. We have tested and compared five feature extraction approaches: gray-map, Sobel, Spatial Gray Level Dependence Matrices (SGLDM), Average Fraction Under the Minimum (AFUM) [11] and a new approach based on the previous one, called Set of Fractions Under the Minimum (SFUM). In all of these approaches, a k-NN classifier based on kd-trees has been used.

2 Approach

In a classical classifier [8], each object for training and test is represented by a feature vector, and a discrimination rule is applied to classify a test vector. In the image classification problem, this feature vector is usually obtained using a pixel-based method (local image properties are computed at each pixel).

In our work, in order to model both cancerous tissue and healthy tissue, the training images are sampled using a sliding window W. A feature vector v_1 is formed by applying one of the approaches mentioned above over the pixels in W. Then, Principal Component Analysis (PCA) is applied to v_1, forming a new vector v_2 of lower dimensionality. This vector v_2 (prototype), together with a label (*cancer/no-cancer*) is stored into the model. During the evaluation phase, test points are classified using a k-nearest neighbors (k-NN) scheme with an approximate search method based on kd-trees [1].

Since tumors present large variations in size, a multi-resolution scheme has been used to overcome the scale-invariant problem. For this purpose, images containing a tumor are properly scaled in the training phase, in order to normalize the lesion to a fixed size. Then, during the test, each image is evaluated at several different scales so that a tumor, if present, can be detected in one of these scales.

2.1 Feature Extraction

Every pixel in a digitized mammogram can be represented by a feature vector computed according to their neighbor pixels. In our case, given a pixel, the neighborhood is defined by a square window W_n of $n \times n$ pixels centered in this pixel. This window is shifted along the image and processed to extract the features of every pixel in the mammogram. The window W_n is preprocessed by normalizing the gray-level values of its pixels, so as the mean is zero and the standard deviation is one. Then, this window is processed in different ways, according to different feature extraction approaches: gray-map, Sobel Filter, SGLDM, AFUM [11] and SFUM. All these methods, but the AFUM, obtain feature vectors whose dimensionality depends on the value of the parameter n. The dimensionality of the feature vectors are reduced by using principal component analysis (PCA). The AFUM algorithm, instead of generating feature vectors, gives a single feature for each pixel, which is be used directly as a confidence value to classify it. Each of the feature extraction methods proposed in this work are subsequently discussed:

Gray-Map. Although using the gray-map around the pixel as its feature vector is a very simple option, we have found that gray-maps followed by PCA, extracted massively for every pixel neighborhood in the image and stored as prototypes, gives very consistent results compared to other more sophisticated feature extraction methods usually employed in image recognition [15, 16].

To evaluate this approach, n^2-dimensional vectors are formed with the normalized gray-level values of the window W_n. Then, these vectors are projected into a lower-dimensionality space by means of PCA and stored as prototypes.

Sobel. A common way to enhance objects edges in an image is by convolution with a Sobel Filter. The Sobel filter consists of two kernels which detect horizontal and vertical changes in an image. If both kernels are applied to an image, the results can by used to compute the magnitude and direction of the edges in the image. In our approach we only compute the magnitude of the edges.

This method has the advantage of short computation time. Nevertheless it is strongly vulnerable to noise and use to produce disconnected contours. It is implicitly tuned to detect edges three pixels wide.

SGLDM. Many approaches to tissue segmentation in medical images are based on texture descriptors such as Spatial Gray Level Dependence Matrices (SGLDM, also known as Coocurrence Matrices) [14], Fractal features [5], and other kinds of textural features [12]. In this work, we have tested SGLDM descriptors.

We have used a subset of the SGLDM descriptors defined in the original work by Haralick et al. [10]: Angular Second Moment, Contrast, Correlation, Variance, Inverse Difference Moment, Sum Average, Sum Variance, Sum Entropy, Entropy, Difference Variance and Difference Entropy.

To compute the SGLDM descriptors, each image had previously been preprocessed using standard vector quantization so that its number of gray-levels was reduced from 256 to exactly 20. The 11 descriptors cited above were extracted from 16 matrices corresponding to four angles (0, $\frac{\pi}{4}$, $\frac{\pi}{2}$ and $\frac{3\pi}{4}$) and 4 distances (1 to 4 pixels), giving a total of 176 parameters. These parameters were normalized using a basis computed from a subset of the training set, giving a variance of approximately 1.0. Finally, each resulting 176-dimensional vector was projected into a lower-dimensionality space by means of PCA.

AFUM. In [11], Heath and Bowyer define a mass detection algorithm called average fraction under the minimum (AFUM). Given a pixel p_{ij}, the fraction of pixels at a distance r_2 from p_{ij} that have a lower intensity value than all the pixels in a circle defined by center (i, j) and radius r_1, is computed. This calculation is done over a range of r_1 and r_2, and the average of all these values is computed. The advantage of this algorithm is that it is invariant to rotations and it does not require a learning or parameter training process.

SFUM. Although it can be and advantage to have a single feature for each pixel, as in the AFUM algorithm described above, a loss of information occurs when the average of a set of features is computed. To try to avoid it, we propose a variant of the AFUM algorithm. Instead of computing the average, the whole set of fractions under the minimum (SFUM) is taken into account, giving a feature vector instead of a single scalar. If the dimensionality of this vector is considered too high, it can be reduced by means of PCA. The feature vectors computed for each window W_n are stored as prototypes. Finally, a test point is classified according to the k-NN rule described in Section 2.3.

2.2 Building the Model

For all the methods presented in Section 2.1 but the AFUM, a model with a number of prototypes of both classes (*cancer* and *non-cancer*), extracted from 297 different mammograms, was built.

We have a representative sample of class *cancer*, but we do not have a representative sample of healthy tissue. A technique often used to avoid the problem of using a huge training set for healthy tissue is bootstrapping, selectively adding images to the training set as the classifier is trained [9].

To do that, an initial model was first built from the central pixels of all the regions labeled with some type of tumor, and from the pixels of the 10% of healthy mammograms in the dataset. Each region containing a tumor was scaled to a fixed size of 20×20 pixels and rotated every 6 degrees to obtain a rotational invariant representation. These rotations were not performed in the case of AFUM and SFUM algorithms, as these approaches are inherently invariant to rotations.

Secondly, the remaining non-cancerous mammograms that were not employed in the previous step, were used to expand the training set using the aforementioned bootstrapping scheme. In each bootstrap iteration, feature vectors of one different mammogram were extracted and classified using the model built in the last iteration (or the initial model for the first iteration). Only those vectors classified as *cancer* (false positives) were added to the model.

2.3 *k*-NN Classification

The k-Nearest Neighbor (k-NN) rule is a powerful and robust classification method with well-known theoretical properties and ample experimental success in many pattern recognition tasks. The major drawback often cited for this method is the computational cost. In this work, approximate nearest neighbors search in *kd*-trees has been used to reduce the temporal cost [1].

Kd-trees and the approximate search algorithm have been successfully used in other pattern recognition tasks, allowing to drastically decrease the search time while keeping recognition rates [17].

In this work, to classify a test point, instead of classifying it accordingly to the most voted class, a confidence criterion function that depends on the distances to the k-nearest neighbors has been used:

$$f_c = \frac{\displaystyle\sum_{i \in s_c} \frac{1}{d(p, n_i)}}{\displaystyle\sum_{i=1}^{k} \frac{1}{d(p, n_i)}},$$

where f_c is the confidence of class c, $d(p, n_i)$ is the distance from the test point p to the $i-$neighbor and s_c is the set of sub-indices of the prototypes belonging to class c among the k nearest neighbors found $n_1 \dots n_k$. Since ours is a two-class problem, only the confidence of one class, *cancer*, is computed. When a threshold T is set, a pixel is considered cancer if that confidence is larger than T. In our case, instead of selecting a specific threshold, classification rates are computed for different values of T, generating a receiver operating characteristic (ROC) curve.

To avoid using the same images (and therefore the same prototypes) in training and test, a leave-one-out technique has been employed, guaranteeing that when a mammogram is being evaluated, not only that one, but the other mammogram of the same patient, are left out from the training set.

3 Experiments and Results

The proposed approaches were applied to the Mammographic Image Analysis Society (MIAS) Database [18]. The MIAS database contains 322 mammograms corresponding to the left and right breasts of 161 patients. From the 322 mammograms, 115 have some kind of abnormality and the remaining 207 belong to patients without any tumor. Abnormalities are classified in 6 different groups: calcifications, circumscribed masses, spiculated masses, architectural distortions, asymmetries and other ill-defined masses. Unfortunately, in the case of calcifications, the ground truth region contains much more healthy tissue than affected tissue. This is due to the shape of calcifications, which are small lesions spreaded into a wide area. For this reason, calcifications were not included in the training set nor evaluated in the test. Leaving the calcifications out, the database consists of 90 cancerous mammograms of a total of 297. Figure 1 shows a mammogram with a spiculated tumor. It can be appreciated how the region with the affected tissue does not differ much from other regions corresponding to healthy tissues.

A number of experiments were carried out to compare the performance of the feature extraction methods proposed. An evaluation set consisting of 3398 regions of interest (ROIs) was created. Of the 3398 regions, 90 correspond to the central pixel of each tumor defined in the dataset, 820 correspond to manually

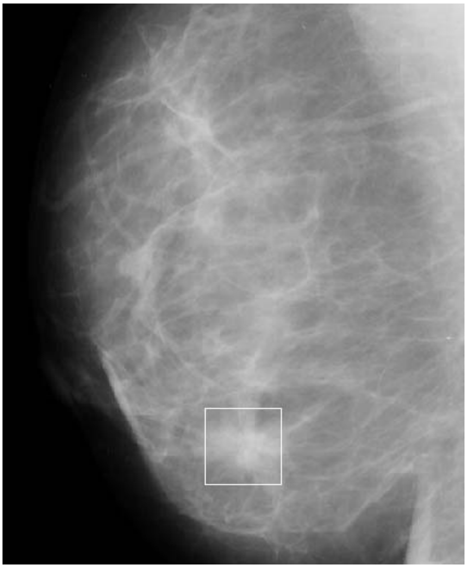


Fig. 1. Mammogram with spiculated cancer.

selected pixels of different mammograms with a high similarity to cancerous tissue, and the remaining 2488 correspond to healthy tissue randomly selected.

Each of the methods proposed in Section 2.1 was evaluated using the test set described above. For each ROI, feature vectors (or a single feature in the case of AFUM algorithm) were computed at different scales. These feature vectors (test points) were classified as explained in Section 2.3, giving a confidence value for each tested scale. In the case of the AFUM algorithm, the value of the feature computed was used directly as confidence value. The mean of the confidence values computed at each scale was then obtained, giving a single confidence value for each ROI.

The true positive rate and false positive rate of the evaluated regions were computed at different thresholds. These rates were represented by a response receiver operating characteristic (ROC) curve. Figure 2 shows the ROC curves obtained in each of the feature extraction methods proposed, while Table 1 shows the areas under the ROC curves (AUC). The parameters used in these experiments were: 30 nearest neighbors, 30 principal components, 11 scales, and a sliding window W_n of 16×16.

Table 1. Area under the ROC curve (AUC) obtained in each of the feature extraction methods proposed.

Method	Gray Map	SGLDM	Sobel	AFUM	SFUM
AUC	0.911	0.814	0.807	0.854	0.910

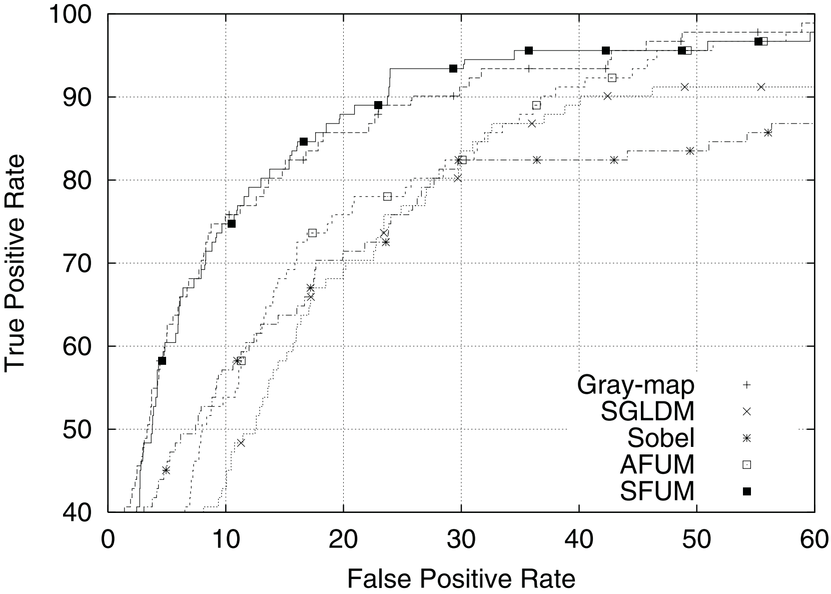


Fig. 2. ROC curve obtained in each of the feature extraction methods proposed.

4 Conclusion

This work presents a comparison of five different feature extraction techniques for breast cancer detection. The results obtained show that the Gray-map and SFUM methods are clearly better than the rest. The combination of a straightforward feature extraction method as the Gray-map and the Principal Components Analysis seems to be a good approach for this problem. Also, it is remarkable that the AFUM algorithm, being a simpler unsupervised method, achieve a good behavior for this task. The proposed SFUM method improves the behavior of the original AFUM as we had supposed. More complicated feature extraction methods like SGLDM and edge-based method like Sobel filter convolution seem not to work properly for this task.

References

1. S. Arya et al. An optimal algorithm for approximate nearest neighbor searching. *Journal of the ACM*, 45:891–923, 1998.
2. K.J. Bovis and S. Singh. Detection of masses in mammograms using texture features. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 2, pages 2267–2270, 2000.
3. Guido M. Te Brake and Nico Karssemeijer. Comparison of three mass detection methods. In *Digital Mammography*, pages 119–126, Dordrecht, The Netherlands, 1998. Kluwer Academic Publishers.

4. R. Campanini and A. Bazzani et al. A novel approach to mass detection in digital mammography based on support vector machines (svm). In *Proceedings of the 6th International Workshop on Digital Mammography (IWDM)*, pages 399–401, Bremen, Germany, 2002. Springer-Verlang.
5. C. Chen, J. Daponte, and M. Fox. Fractal feature analysis and classification in medical imaging, 1989.
6. I. Christoyianni and A. Koutras et al. Computer aided diagnosis of breast cancer in digitized mammograms. *Computerized Medical Imaging and Graphics*, 26:309–319, 2002.
7. Wan Mimi Diyana, Julie Larcher, and Rosli Besar. A comparison of clustered microcalcifications automated detection methods in digital mammogram. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 385–388, Hong Kong, 2003.
8. R. Duda and P. Hart. *Pattern Recognition and Scene Analysis*. John Wiley, New York, 1973.
9. H.A. Rowley H.A. and S. Baluja et al. Neural network-based face detection. *IEEE Transactions on PAMI*, 20(1):23–38, 1998.
10. R M Haralick et al. Textural features for image classification. *IEEE Trans. SMC*, 3(6):610–621, 1973.
11. Michael D. Heath and Kevin W. Bowyer. Mass detection by relative image intensity. In *Proceedings of the 5th International Workshop on Digital Mammography (IWDM-2000)*, pages 219–225. Medical Physics Publishing, 2000.
12. M. Insana et al. Analysis of ultrasound image texture via generalized rician statistics. *Opt. Eng.*, 25:743–748, 1986.
13. M.J. Lado and P.G. Tahoces et al. Evaluation of an automated wavelet-based system dedicated to the detection of clustered microcalcifications in digital mammograms. *Med. Inform. In Med.*, 26:149–163, 2001.
14. G. Landeweerd and E. Gelsema. The use of nuclear texture parameters in the automatic analysis of leukocytes. *Pattern Recognition*, 10:57–61, 1978.
15. R. Llobet and J. C. Perez-Cortes. Breast cancer detection in digitized mammograms using non-parametric methods. In *Proceedings of the 2nd International Conference on Advances in Biomedical Signal and Information Processing (MEDSIP)*, volume 1, pages 281–287, Sliema, Malta, 2004.
16. R. Llobet, A.H. Toselli, J.C. Perez-Cortes, and Alfons Juan. Computer-aided prostate cancer detection in ultrasonographic images. In *Proceedings of the 1st Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*, volume 1, pages 411–419, Puerto de Andratx (Mallorca, Spain), 2003.
17. J.C. Perez-Cortes, J. Arlandis, and R. Llobet. Fast and accurate handwritten character recognition using approximate nearest neighbours search on large databases. In *Workshop on Statistical Pattern Recognition SPR-2000*, volume 1876 of *Lecture Notes in Artificial Intelligence*, pages 767–776, Alicante (Spain), 2000.
18. J. Suckling and J. Parker et al. The mammographic images analysis society digital mammogram database. In *Exerpta Medica. International Congress Series*, volume 1069, pages 375–378, 1994.
19. G.M. te Brake and N. Karssemeijer. Automated detection of breast carcinomas that were not detected in a screening program. *Radiology*, 207:465–471, 1998.
20. M. Wallis and M. Walsh et al. A review of false negative mammography in a symptomatic population. *Clin Radiol*, 44:13–15, 1991.

Part VII

Biometrics

Multiscale Approach for Thinning Ridges of Fingerprint

Xinge You[1,2], Bin Fang[1], Yuan Yan Tang[1], and Jian Huang[1]

[1] Department of Computer Science, Hong Kong Baptist University
{xyou,bfang,yytang,jhuang}@comp.hkbu.edu.hk
[2] Faculty of Mathematics and Computer Science, Hubei University, P.R. China

Abstract. This paper presents a robust multiscale method to create thinned ridge map of fingerprint for automatic recognition by employing an elaborately designed wavelet function. Properties of the new wavelet function are substantially investigated. Some desirable characteristics of the local minimum produced by wavelet transform show that they are suitable to describe skeleton of ribbon-shape objects. A multiscale thinning algorithm based on the modulus minima of wavelet transform is proposed. The proposed algorithm is able to improve the skeleton representation of the ridge of fingerprint without side-effects and limitations of previous methods. Thinned ridge map helps to facilitate minutiae extraction for matching. Experiments validated effectiveness and efficiency of the proposed method.

1 Introduction

Among various biometric techniques, fingerprint recognition is regarded as the most popular and reliable means for automatic personal identification. During past years, Automatic Fingerprint Identification System (AFIS) has received increasingly more attention [1–3]. AFISs dealing with small size database have reached a satisfactory level, the trend of research is to develop algorithms for fingerprint identification systems that can search relatively large databases for online applications.

Most classical fingerprint recognition algorithms [1–4] take the minutiae as the distinctive features to represent the fingerprint in the matching process. Minutiae are ridge ending and ridge bifurcations, as shown in Fig. 1. Minutia extraction is a key step in building a high performance AFIS. It mainly consists of following steps: orientation field estimation, ridge thinning and minutiae extraction. Generally speaking, if an ideal thinned ridge map can be obtained, then minutia extraction is just a trivial task of identifying singular points. Thus the estimation of thinned ridge map plays an important role in AFIS. Therefore, an algorithm to extract ridge map accurately and robustly is desired [1–4].

Many algorithms have been proposed for the extraction of the thinned ridge map, but the results are not satisfactory, especially for poor quality fingerprint images. Due to imperfections of the thinned ridge map, minutiae extraction methods are prone to missing some real minutiae while picking up spurious

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 505–512, 2005.

points (artifacts) [1, 3–5]. Imperfections of the ridge map can also cause errors in determining the location coordinates of the true minutiae and their relative orientation in the image.

Thinned algorithms can be classified into direct and indirect computing methods [6, 7]. Some drawbacks of the direct computing methods badly affect their performance. The generated skeletons are in discrete forms which is not helpful for recognizing the underlying shape. Secondly, even if skeleton pixels are linked, the resulting skeleton may not be centred inside the underlying shape due to the use of discrete data.

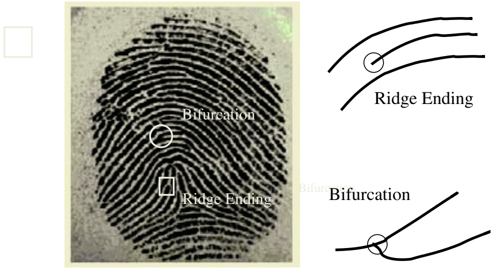


Fig. 1. A minutiae feature (a ridge ending and a bifurcation) in a fingerprint image.

Alternatively, skeletons can be computed indirectly [7, 8]. In an indirect skeletonization processing, the skeleton of shape is referred to as the locus of the symmetric points or symmetric axes of the local symmetries of the shape contour [9]. Different local symmetry analyse maybe result in different symmetric points, and hence different skeletons are produced, such as *Symmetric Axis Transform* [9], *Smoothed Local Symmetry* [10], *Process-Inferring Symmetry Analysis* [11] and the latest maximum modulus symmetry of wavelet transform [8]. However, the major problem of the indirect computing technique lies in the difficulty of accurately identifying local symmetries of the underlying shape.

After all, the computation complexity of most methods is high. It usually takes $O(n^3)$ [7] time to compute the skeleton of an $n \times n$ image by thinning. Obviously, time-consuming thinned methods are not able to meet requirements of on-line verification application involving large databases.

In this paper, we present a novel algorithm for thinning ridge of the fingerprint based on the local minima of wavelet transform moduli. This proposed thinned algorithm improves the quality of thinned ridge structures and make thinned results be more suitable for minutiae extraction than previous methods. The skeleton computing is implemented directly over a process depending on edges and contours detected.

This paper is organized as follows: First, in Section 2, Some properties of the local minima of wavelet transform moduli with a new wavelet function are discussed. In Section 3, a multiscale algorithm based on wavelet transform for extracting thinned ridge map of fingerprint is proposed. Experimental results are provided in Section 4, with conclusion remarks presented in Section 5.

2 Minima of New Wavelet for Images

Edge points of shape in the image often locate where the image intensity has sharp transition. The local maximum of the absolute value of the first derivative are sharp variation points of $f * \theta_s(x)$. Where a real smoothing function $\theta(x)$ satisfiies $\theta(x) = O(\frac{1}{(1+x^2)})$ and whose intergral is nonzero. It can be viewed as the impulse response of a low-pass filter. Let $\theta_s(x) = (\frac{1}{s})\theta(\frac{x}{s})$ and$f(x) \in L^2(R))$. Singular points (such as edges in images) at the scale s are defined as local sharp variation points of $f(x)$ smoothed by $\theta_s(x)$. Whereas the skeleton point of underlying shape should be midpoint between the two edge points along the gradient and where the image intensity of shape has the slowest transition. Hence the skeleton points of the underlying shape correspond to the local minima of the wavelet transform modulus $|\nabla Wf(s,x)|$ which is called wavelet minima. From a viewpoint of mathematical analysis, there should be a local minimum locating between the two consecutive local maxima of $|\nabla Wf(s,x)|$. Moreover, if the wavelet has a compact support, a value of $Wf(s,x_0)$ depends upon the values of $f(x)$ in a neighborhood of x_0, of size proportional to the scale s. Thus, for the fixed scale s and some "fine" wavelet, , the wavelet minima point may locate at the center between two consecutive modulus maxima points and is independent of the scale. Further all these minima points form the skeleton of the underlying shape. Hence, the skeleton of the underlying shape in an gray image can be measured by computing wavelet maxima. A typical multiscale analysis of image is implemented by smoothing the surface with a convolution kernel $\theta(x,y)$ that is dilated at dyadic scales s. Such an edge detector can be computed with two wavelets that are the partial derivative of $\theta(x,y)$

$$\psi^1(x,y) := \frac{\partial}{\partial x}\theta(x,y) \quad \text{and} \quad \psi^2(x,y) := \frac{\partial}{\partial y}\theta(x,y).$$

Let us denote $\theta_s(x,y) := \frac{1}{s^2}\theta(\frac{x}{s},\frac{y}{s})$. For wavelets $\psi^i(x,y),\ i = 1,2$, defined above, their scale wavelet transforms can be written as

$$W_s^1 f(x,y) = (f * \psi_s^1)(x,y) = s\frac{\partial}{\partial x}(f * \theta_s)(x,y), \tag{1}$$

$$W_s^2 f(x,y) = (f * \psi_s^2)(x,y) = s\frac{\partial}{\partial y}(f * \theta_s)(x,y). \tag{2}$$

As far as the skeleton extraction is concerned, the desired wavelet not only detects the edge points by the local maxima of the wavelet transform moduli, but also describes the position of the center of the shape by using a local maxima. Although lots of wavelet functions have been constructed, the construction of an appropriate wavelet for such particular application is still a great challenge.

We note that the most of multiscale shape representation or analysis methods are based on Gaussian kernael. However, as pointed out by Mokhtarian and Mackworth [12], the representation should satisfy the efficiency and ease of implementation in order that it is useful for practical shape recognition tasks in computer vision. Obviously, Gaussian kernel does not satisfy such criteria since the computational burden is high at large scales and it is not compact support as well.

The B-spline is a good approximation to the Guassian function with inheriting almost all good properties of the Gaussian funnction [13]. It has been shown that the B-spline wavelet performs better than other wavelets for singularities detection [13]. B-spine derived scale-space exhibits many advantages. For example, it provides fast algorithms and parallel implementation at multiple scales. Unfortunately, it has been proved that it fails to process the Dirac-structure edge due to be not width-invariant of its wavelet maxima [14]. Consequently, they do not satisfy the such extra requirements for skeleton extraction [14, 15].

The new wavelet, $\psi(x, y)$, which is constructed in our latest work [8, 14], not only has the advantages of the Gaussian function and quadratic spline, but also it possesses some desirable properties of its maxima. In this paper, we further proved mathematically the following properties of its modulus minima. These properties can be summarized the following theorem:

Theorem 1. *Let l_d be a straight segment of the ridge of fingerprint with width d. For different scales of wavelet transform with the new wavelet, the wavelet maxima locate in the mid position of the ridge segment and include the inherent skeleton of the ridge segment. Especially, if the scale is much larger than the width of the ridge, the local minimum points exist uniquely and form skeleton lines of the ridge segments which consist of a series of single pixel. Further, the two contours obtained from the wavelet minima are symmetric with respect to this local skeleton line along the gradient direction. Otherwise, all local minimum points form the thinned ribbon consisting of multiple pixels in the mid of ridge segment.*

3 Multiscale Algorithm for Thining Ridge Map

The above properties imply that the locations of wavelet minima cover exactly the inherent central line of the shape. Meanwhile, the location of maxima of the wavelet transform moduli, which locate nearly at the points of the original boundary, form the two new lines and they are symmetrical with respect to the inherent central line of the ridge which can be identified by the local minima points. Therefore the connective curve of all minimum points of WT modulus is defined as primary skeleton of the ridge. Thus a simple and direct strategy for extracting the skeleton of ridge of fingerprint is to compute the wavelet minima. In practice, the detection of the wavelet minima in the discrete domain can be implemented analogously as the local maxima of the WT moduli in [8]. Ideally, the skeletons of a ridge of fingerprint are represented by a set of thin curves which consist of single pixels rather than by a ribbon-shape objects which consists of multiple pixels. It is shown in the theorem 1 that if and only if the scale of wavelet transform matches well with the width of the shape, namely, its value is much bigger than the width of the shape, the modulus minimum point between two homologous modulus maximum points exists uniquely. Otherwise, maybe there exists numerous modulus minimum points and all of these points form a continuous region or bandwidth, which is called skeleton ribbon or primary skeleton. In practice, it is difficult to choose the suitable scale of the WT according to the width of shape structure so that the skeleton locus obtained from the wavelet minima consists of the single pixels. For most cases, the primary skeleton obtained from the modulus-minimum-based algorithm is generally the bandwidth skeleton consisting of multiple pixels than the perfect skeleton line containing a single pixel. Even if the relatively big scale is favorable to process on wide-structure shape and may result in a single-pixel skeleton, but it usually suffers from too heavy computational cost.

To solve this problem, the following mulitiscale-based approach is proposed. Our basic idea is as follows: For each input image, we randomly choose a scale of wavelet transform and extract its corresponding skeleton of the underlying ridge of fingerprint by computing all wavelet minima. Hence, all these local minimum points produce the primary skeleton ribbon of the underlying ridge which consists of multiple pixels. Obviously, these primary skeleton ribbons are apparently thinner than the original shapes and preserve exactly the topological properties of the original ridge. Likewise, we choose a much smaller scale than the prior one to perform the second wavelet transform on the image which contains generated skeleton ribbons, and compute the second level skeletons. The above procedure is iterated until the central curves which consist of single pixels is eventually extracted from the underlying ridge of fingerprint. This procedure is called Multiscale Thinning Algorithm (*MTA*) and designed as follows:

Algorithm 1 *(Multiscale Thinning Algorithm)*

Step 1 *Input an fingerprint image $f(x,y)$. Evaluate a threshold T_g for the gray value of the original image processed for further partitioning the underlying ridge from the background (if necessary);*

Step 2 *Select randomly an initial scale to perform WT with the wavelets defined by Eq. (1, 2) on the input image.*

Step 3 *Calculate the wavelet minima and set an initial threshold T_{noise} for the modulus of WT.*

Step 4 *Compute the primary skeleton points by using the thresholding, namely, compare $|\nabla W_s f(x,y)|$ with the initial threshold T_{noise} and preserve all points whose modulus are less than an initial threshold T as the candidate points of initial skeleton.*

Step 5 *If the previously obtained skeleton consist of multiple pixels, then choose the second scale and repeat Step 2, otherwise, exit produce the the ridge map from the final skeleton image by smooth processing.*

4 Experiments

In the experiment, which is illustrated in Fig. 2, displays the full procedure of the proposed thinned algorithm step by step. A grey scale image consisting of the distracting background is shown in Fig. 2(a). As mentioned previously, most of existing algorithms often fail to directly process these gray images. The

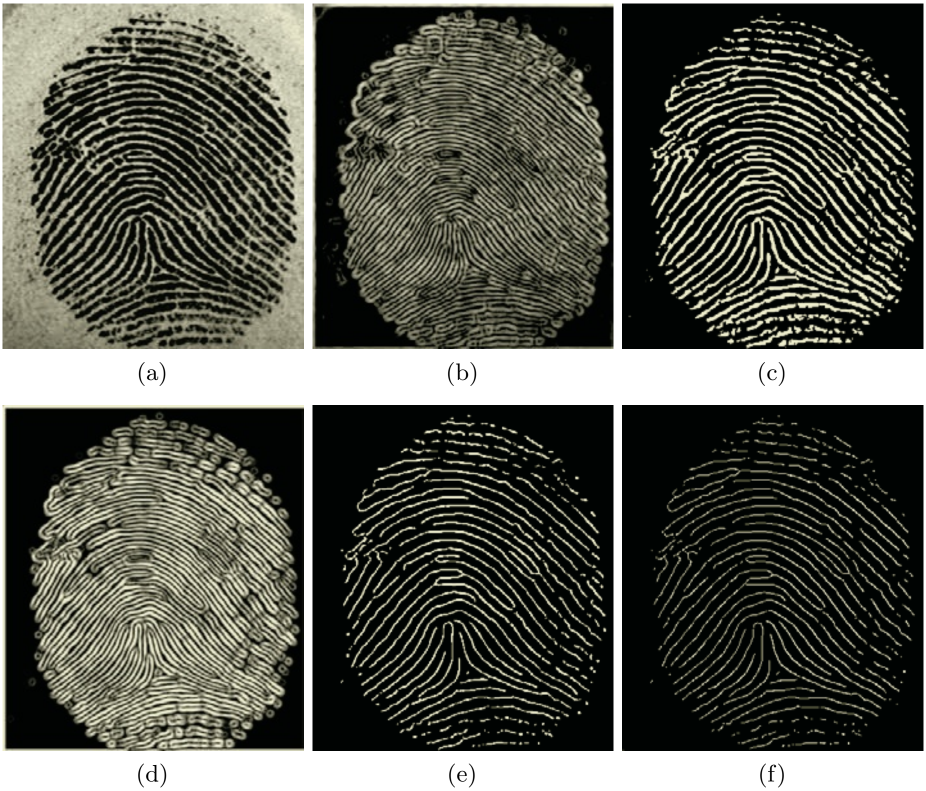


Fig. 2. (a) The original fingerprint image; (b) the raw output of modulus obtained from WT with scale $s = 6$; (c) the primary skeleton ribbon from the first WT; (d) the raw output of modulus obtained from the second WT with scale $s = 4$; (e) the thinned result from the second WT, (f) The final ridge map from the third WT with scale $s = 4$.

modulus image obtained from the WT with scale $s = 6$ is presented in Figs. 2(b) To remove the noise of the modulus image in the distracting background, the modulus image is processed with the threshold $T_noise = 0.31$. Namely the modulus is set to zero if the modulus is smaller than the threshold T_noise. The primary skeleton ribbon from the modulus minima of the first WT is shown in Fig. 2(c). Next, for the primary skeleton ribbon from Fig. 2(c), we perform the second WT with scale $s = 4$. By taking the threshold $T = 0.11$ to compute the modulus minima for the output of moduli of the second WT, the second thinned results are obtained in Fig. 2(e). Obviously, these skeletons consists multiple pixels. Therefore, we further perform the third WT for the thinner skeleton ribbons in Fig. 2(e). The final result is shown in Fig. 2(f) and it meets the requirements for extraction minutiae.

We compare the executed time of the proposed Wavelet-Based method for thinning the ridges with that of Zou's method (ZM) [7], Zhang and Suen's

Table 1. Comparsion of the computational time and matching rate among various algorithms.

Methods	Comption time (seconds)			Matching rate (%)		
	FI1	FI 2	FI 3	FI1	FI 2	FI 3
Proposed Method	21.37	15.25	18.71	98.94	97.63	98.25
ZM	25.25	19.15	24.21	97.51	94.25	96.95
ZSM	28.78	21.73	25.87	94.23	93.55	94.73
PCBA	24.34	20.53	23.49	95.13	92.79	93.57

method (ZSM) [6] , and Principal-curve-based method (PCBA) [16] by testing three fingerprint images which are randomly selected from the database in [1]. Furthermore, their corresponding minutia sets are estimated based on the thinned ridge maps obtained from four different methods. By using matching algorithm and the standard template set of minutia provided in [1], some matching results are shown in Table 1 as well. Obviously, computational speed of skeletonization and accuracy of minutia sets from thinned ridge map based on the proposed algorithm are greatly improved.

5 Conclusions

In this paper, a new wavelet-based method for extracting ridge map from fingerprint images is proposed. It depends on some desirable properties of the constructed wavelet function, and it offers a different view and increase of insight on the problem of computing skeleton of ribbon-shape objects. Some properites of the new wavelet function are analyized mathematically. It is pointed out that the method based on the estimation of the wavelet minima is able to directly extract skeleton of ridge with improved the accuracy. The thinned method can be implemented very efficiently with relatively lower computational time than other traditional methods. These improvements are very helpful for identifying minutia in the fingerprint verification system.

Acknowledgments

This research was partially supported by a grant (60403011 and 10371033) from National Natural Science Foundation of China and a grant (2003ABA012) from Hubei Provincial Science &Technology Department, and a grant (20045006071-17) from Wuhan government, China. This research was also supported by the grants (RGC and FRG) from Hong Kong Baptist University.

References

1. A. Jain, L. Hong, and R. Bolle, "On-line fingerprint verfication." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, pp. 302–314, 1997.

2. A. M. Bazen and S. H. Gerez, “systematic methods for the computation of the directional fields and singular points of fingerprints.” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 905–919, 2002.
3. ——, “Fingerprint matching by thin-plate spline modeling of elastic deformations.” *Pattern Recognition*, vol. 36(8), pp. 1859–1867, 2003.
4. A. Ross, S. C. Dass, and A. Jain, “Estimating fingerprint deformation.” *Proceedings of the international Conference on Biometric Authentication (ICBA).*, vol. 9(5), pp. 846–859, 2004.
5. G. Marcialis and F. Roli, “Perceptron-based fusion of multiple fingerprint matchers,” 2003.
6. L. Lam, S. W. Lee, and C. Y. Suen, “Thinning Methodologies - a Comprehensive Survey”, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. vol. 14, pp. pp. 869–885, 1992.
7. J. J. Zou, “Skeleton represetation of ribbon-like shapes,” PhD thesis, School of Electical and Information Engineering, University of Sydney, Sydney, March 2001.
8. Y. Y. Tang and X. G. You, “Skeletonization of ribbon-like shapes based on a new wavelet function,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1118–1133, 2003.
9. H. Blum, “A transformation for extracting new desxriptors of shape,” W. Wathen-Dunn, Eds., Models for the Perception of Speech and Visual Form, pp. 362-380. Massachusetts: The MIT Press, 1967.
10. M. Brady., “Criteria for Representation of Shape,” J. Beck and B. Hope and A. Rosenfeld, Eds., Human and Machine Vision, pp. 39-84. New York: Academic Press, 1983.
11. M. Leyton, “A process-grammar for shape”, *Artifial Intell.*, vol. vol. 34, pp. pp. 213–247, 1988.
12. F. Mokhtarian and A. K. Mackworth, “A theory of multiscale curvature-based shape representation for planar curves”, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. vol. 14, no. 8, pp. 789–805, 1992.
13. Y.-P. Wang and S. L. Lee, “Scale-space derived from B-splines”, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. vol. 20, no. 10, pp. 1040–1050, 1998.
14. Y. Y. Tang, L. H. Yang, and J. M. Liu, “Characterization of Dirac-Structure Edges with Wavelet Transform”, *IEEE Trans. Systems, Man, Cybernetics (B)*, vol. vol. 30, no. 1, pp. pp. 93–109, 2000.
15. L. H. Yang, X. You, R. M. Haralick, I. T. Phillips, and Y. Tang, “Characterization of Dirac Edge with New Wavelet Transform”, in *Proc. 2th Int. Conf. Wavelets and its Application*, vol. 1, 2001, pp. 872–878.
16. B. Kegl and A. Krzyżak, “Piecewise linear skeletonization using principal curves,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 59–74, 2002.

Discriminative Face Recognition Through Gabor Responses and Sketch Distortion

Daniel González-Jiménez and José Luis Alba-Castro

Departamento de Teoría de la Señal y Comunicaciones
Universidad de Vigo, Spain
{danisub,jalba}@gts.tsc.uvigo.es

Abstract. We present an inherently discriminative approach to face recognition. This is achieved by automatically selecting key points from lines that sketch the face and extracting textural information at these locations. As the distribution of the lines depend on each individual face, the selected points will be person-dependent, achieving discrimination in an early stage of the recognition process. A robust shape matching algorithm has been used for the correspondence problem, and Gabor responses have been extracted at final points so that both shape and textural information are combined to measure similarities between faces. Face verification results are reported over the well known XM2VTS database.

1 Introduction

One of the most successful approaches to automatic face recognition is the Elastic Bunch Graph Matching algorithm (EBGM) [2]. It combines local and global representation of the face by computing multi-scale and multi-orientation Gabor responses (jets) from a set of the so-called fiducial points, located at specific face regions (eyes, tip of the nose, mouth..., i.e. "universal" features). Finding every fiducial point relies on a matching process between the candidate jet and a bunch of jets extracted from the corresponding fiducial points of different faces. This matching problem is solved by maximizing a function that takes texture and geometrical distortion into account. In this way, there are several variables that can affect the accuracy of the final positions, as differences in pose, illumination conditions and insufficient representativeness of the stored bunch of jets. Once fiducial points are adjusted, only textural information (Gabor jets) is used in the classifier.

The main differences between EBGM and our approach are focused on the way we locate and match fiducial points and on the final dissimilarity function that does not only use texture but also geometrical information. Our method locates salientable points in face images by means of the ridges and valleys operator. As only some basic image operations are needed, the computational load is reduced from the original EBGM algorithm and, at the same time, possible discriminative locations are found in an early stage of the recognition process. In this sense we say that this method is inherently discriminative, in contrast to trainable parametric models. The set of selected points turned out to be quite

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 513–520, 2005.

robust against illumination conditions and slight variations in pose. Many of the located fiducial points belong to “universal” features, but many others are person-dependent. So, EBGM locates a pre-defined set of “universal” features and our approach finds a person-dependent set of features. The correspondence between fiducial points of two faces only uses geometrical information and it is based on shape contexts [1]. As a byproduct of the correspondence algorithm, we extract measures of local geometrical distortion. Gabor jets are also calculated from the adjusted points and the final dissimilarity function compiles geometrical and textural information.

The different stages of our method are detailed in the next sections. Section 2 explains the operator used to extract face lines. Grid adjustment and selection of points is described in section 3. Section 4 shows the algorithm used to match points between two faces. The Gabor filters used to extract texture are described in section 5. Section 6 introduces different geometrical terms and the *Sketch Distortion* concept used to measure dissimilarity between faces. Experimental results are given in section 7. Finally, conclusions are drawn and future research lines are discussed in section 8.

2 Extraction of Low-Level Geometrical Descriptors

In this work, shape information has been obtained using the ridges and valleys operator because of its robustness against illumination changes [3]. Moreover, the relevance of valleys in face shape description has been pointed out by some cognitive science works [4]. In this paper, we have used the ridges and valleys obtained by thresholding the so-called multi local level set extrinsic curvature (MLSEC) [5]. The MLSEC operator works here as follows: **i)** computing the normalized gradient vector field of the smoothed image, **ii)** calculating the divergence of this vector field, which is bounded and gives an intuitive measure of valleyness (positive values running from 0 to 2) and ridgeness (negative values from -2 to 0), and **iii)** thresholding the response so that image pixels where the MLSEC response is smaller than -1 are considered ridges, and those pixels larger than 1 are considered valleys.

Once the feature descriptor has been properly defined, we have a way of describing fiducial points in terms of positions where the geometrical image features have been detected. For this shape descriptor to be useful in face recognition or authentication, local texture information must be also taken into account. Gabor wavelets are biologically motivated convolution kernels that capture this kind of information and are also quite invariant to the local mean brightness, so an efficient face encoding approach will be to extract texture from these geometrically salience regions.

3 Grid Adjustment

Once the ridges and valleys in a new image have been extracted, we must sample these lines in order to keep a set of points for further processing. There are some

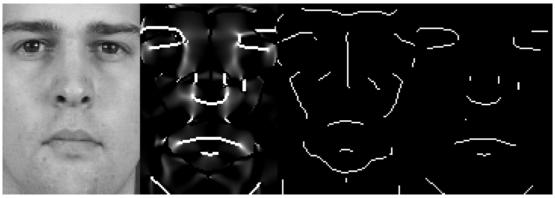


Fig. 1. Left: Original Image. Center-left: Valleys and ridges image. Center-right: Thresholded ridges image. Right: Thresholded valleys image.

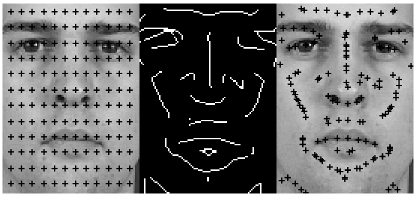


Fig. 2. Left: Original rectangular dense grid. Center: Valleys and ridges sketch. Right: Grid adjusted to the sketch.

possible combinations, in terms of using just ridges, just valleys or both of them, so we will refer to the binary image, obtained as a result of the previous step, as the sketch from now on.

In order to select a set of points from the original sketch, a dense rectangular grid ($\mathcal{N}_x \times \mathcal{N}_y$ nodes) is applied onto the face image and each grid node changes its position until it finds the nearest line of the sketch. So, finally, we get a vector of points $\mathcal{P} = \{\boldsymbol{p_1}, \boldsymbol{p_2}, \ldots, \boldsymbol{p_n}\}$ [1], where $\boldsymbol{p_i} \in \mathbb{R}^2$. These points sample the original sketch, as it can be seen in figure 2.

Scale Normalization. Suppose an incoming face image $\mathcal{F}$, with a constellation of points distributed over the face region: $\mathcal{P} = \{\boldsymbol{p_1}, \boldsymbol{p_2}, \ldots, \boldsymbol{p_n}\}$. To achieve scale invariance, we must measure how big the face is. One way to do this is by adding the distances bettween all points in the constellation, i.e. to proceed as follows:

$$D_{face} = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \|\boldsymbol{p_i} - \boldsymbol{p_j}\| \tag{1}$$

This distance D_{face} gives an idea of the size of the face, so that it can be normalized to standard scale just by resizing the input image by a factor of $r = \frac{D_{std}}{D_{face}}$, where D_{std} is the distance D_{face} for a standard face size. Also, if we are looking for a more accurate estimation of the size of the face, an iterative process can be applied until the ratio $r \in (1-\tau, 1+\tau)$ for a given threshold $\tau > 0$.

[1] $n = \mathcal{N}_x \times \mathcal{N}_y$. Typical sizes for n are 100 or more nodes.

4 Point Matching

Once we have the two face images, $\mathcal{F}_1$ and $\mathcal{F}_2$, at common size, we want to proceed to compute similarity between them. Let $\mathcal{P} = \{\boldsymbol{p_1}, \boldsymbol{p_2}, \ldots, \boldsymbol{p_n}\}$ be the set of points for $\mathcal{F}_1$, and $\mathcal{Q} = \{\boldsymbol{q_1}, \boldsymbol{q_2}, \ldots, \boldsymbol{q_n}\}$ the set of points for $\mathcal{F}_2$.

In order to compare feature vectors extracted at these positions, we must first compute the matching between points from both images. We have adopted the idea described in [1]. For each point i in the constellation, we compute a 2-D histogram h_i of the relative position of the remaining points, so that a vector of distances $\mathcal{D} = \{d_{i1}, d_{i2}, \ldots, d_{in}\}$ and a vector of angles $\boldsymbol{\theta} = \{\theta_{i1}, \theta_{i2}, \ldots, \theta_{in}\}$ are calculated for each point. As in [1], we employ bins that are uniform in log-polar space, i.e. the logarithm of distances is computed. Each pair $(\log d_{ij}, \theta_{ij})$ will increase the number of counts in the adequate bin of the histogram.

Once the sets of histograms are computed for both faces, we must match each point in the first set $\mathcal{P}$ with a point from the second set $\mathcal{Q}$. A point $\boldsymbol{p}$ from $\mathcal{P}$ is matched to a point $\boldsymbol{q}$ from $\mathcal{Q}$ if the term C_{pq}, defined as:

$$C_{pq} = \sum_k \frac{\left[h_p(k) - h_q(k)\right]^2}{h_p(k) + h_q(k)} \tag{2}$$

is minimized[2]. Finally, we have a correspondence between points defined by ξ:

$$\xi(i) : \boldsymbol{p_i} \Longrightarrow \boldsymbol{q_{\xi(i)}} \tag{3}$$

where $\boldsymbol{p_i} \in \mathcal{P}$ and $\boldsymbol{q_{\xi(i)}} \in \mathcal{Q}$.

Dealing with Rotations in Plane. The vectors of angles $\boldsymbol{\theta} = \{\theta_{i1}, \theta_{i2}, \ldots, \theta_{in}\}$ are calculated taking the x-axis (the vector $(1,0)^T$) as reference. This is enough if we are sure that the faces are in an upright position. But, to deal with rotations in plane, i.e. if we do not know the rotation angle of the heads, we must take a relative reference for the shape matching algorithm to perform correctly. Consider, for the set of points $\mathcal{P} = \{\boldsymbol{p_1}, \boldsymbol{p_2}, \ldots, \boldsymbol{p_n}\}$, the centroid of the constellation $\boldsymbol{c_{\mathcal{P}}}$:

$$\boldsymbol{c_{\mathcal{P}}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{p_i} \tag{4}$$

For each point $\boldsymbol{p_i}$, we will use the vector $\overrightarrow{p_i c_{\mathcal{P}}} = \boldsymbol{c_{\mathcal{P}}} - \boldsymbol{p_i}$ as the x-axis, so that rotation invariance is achieved. Also, the angle between the two images, φ, can be computed as follows:

$$\varphi = \frac{1}{n} \sum_{i=1}^{n} \angle \left(\overrightarrow{p_i c_{\mathcal{P}}}, \overrightarrow{q_{\xi(i)} c_{\mathcal{Q}}}\right) \tag{5}$$

so that the system is able to put both images in a common position for further comparison. If we do not take this angle into account, textural extraction will not be useful for our purposes.

[2] k in (2) runs over the number of bins in the 2D histogram.

5 Extracting Textural Information

The system uses a set of 40 Gabor filters, with the same configuration employed in [2]. These filters are convolution kernels in the shape of plane waves restricted by a Gaussian envelope, as it is shown next:

$$\psi_m\left(\overrightarrow{x}\right) = \frac{\left\|\overrightarrow{k}_m\right\|^2}{\sigma^2} \exp\left(-\frac{\left\|\overrightarrow{k}_m\right\|^2 \left\|\overrightarrow{x}\right\|^2}{2\sigma^2}\right)\left[\exp\left(i\overrightarrow{k}_m \cdot \overrightarrow{x}\right) - \exp\left(-\frac{\sigma^2}{2}\right)\right] \quad (6)$$

where $\overrightarrow{k}_m$ contains information about frequency and orientation of the filters, $\overrightarrow{x} = (x, y)^T$ and $\sigma = 2\pi$.

The region surrounding a pixel in the image is encoded by the convolution of the image patch with these filters, and the set of responses is called a jet, $\mathcal{J}$. So, a jet is a vector with 40 coefficients, and it provides information about a specific region of the image. Each coefficient, $\mathcal{J}_k$, can be expressed as follows:

$$\mathcal{J}_k\left(\mathcal{I}\left(x_0, y_0\right)\right) = \sum_x \sum_y \mathcal{I}(x, y)\psi_k\left(x_0 - x, y_0 - y\right) \quad (7)$$

In the previous step, we have selected n points from the face image, but in order to avoid overlapping between responses of filters and to reduce computational time, we must leave just a few of them, from which we will extract textural information. So, we decided to establish a minimum distance D between each pair of nodes, so that all final positions are separated at least by D. As a consequence, the number of final points, n_D, will be less or equal than n. Let $\mathcal{P}' = \left\{\boldsymbol{p'_1}, \boldsymbol{p'_2}, \ldots, \boldsymbol{p'_{n_D}}\right\}$ denote the set of final points for textural extraction, and let $\mathcal{R} = \left\{\mathcal{J}_{\boldsymbol{p'_1}}, \mathcal{J}_{\boldsymbol{p'_2}}, \ldots, \mathcal{J}_{\boldsymbol{p'_{n_D}}}\right\}$ be the set of jets calculated for one face. The similarity function between two faces, $\mathcal{S}_{\mathcal{J}}\left(\mathcal{F}_1, \mathcal{F}_2\right)$ results in:

$$\mathcal{S}_{\mathcal{J}}\left(\mathcal{F}_1, \mathcal{F}_2\right) \equiv \mathcal{S}_{\mathcal{J}}\left(\mathcal{R}^1, \mathcal{R}^2\right) = \frac{1}{n_D} \sum_{i=1}^{n_D} < \mathcal{R}_i^1, \mathcal{R}_{\xi(i)}^2 > \quad (8)$$

where $< \mathcal{R}_i^1, \mathcal{R}_{\xi(i)}^2 >$ represents the normalized dot product between the i-th jet from $\mathcal{R}^1$ and the correspondent jet from $\mathcal{R}^2$, but taking into account that only the moduli of jet coefficients are used.

6 Measuring Geometrical Deformations: The Sketch Distortion

In [2], no metric information was used in the final similarity measurement between faces. Our system takes into account geometrical deformation to perform authentication. In the previous section, we discarded some of the points for textural extraction, but the complete set of points is used here to measure geometrical distortions. We have defined two different terms:

$$\mathcal{GD}_1\left(\mathcal{F}_1,\mathcal{F}_2\right) \equiv \mathcal{GD}_1\left(\mathcal{P},\mathcal{Q}\right) = \sum_{i=1}^{n} C_{i\xi(i)} \tag{9}$$

$$\mathcal{GD}_2\left(\mathcal{F}_1,\mathcal{F}_2\right) \equiv \mathcal{GD}_2\left(\mathcal{P},\mathcal{Q}\right) = \sum_{i=1}^{n} \left\| \overrightarrow{p_i c_{\mathcal{P}}} - \overrightarrow{q_{\xi(i)} c_{\mathcal{Q}}} \right\| \tag{10}$$

Equation (9) computes geometrical distortion by adding all the individual costs represented in (2). On the other hand, (10) calculates metric deformation by summing the norm of the difference vector between matched points[3].

Geometrical distortions should be large for faces of different subjects and small for faces representing the same person, so that we can think of combining jet dissimilarity, $\left[1 - \mathcal{S}_{\mathcal{J}}\left(\mathcal{F}_1,\mathcal{F}_2\right)\right]$, with weighted metric deformations, resulting in the final dissimilarity function $\mathcal{DS}\left(\mathcal{F}_1,\mathcal{F}_2\right)$:

$$\mathcal{DS}\left(\mathcal{F}_1,\mathcal{F}_2\right) = \left[1 - \mathcal{S}_{\mathcal{J}}\left(\mathcal{F}_1,\mathcal{F}_2\right)\right] + \lambda_1 \cdot \mathcal{GD}_1\left(\mathcal{F}_1,\mathcal{F}_2\right) + \lambda_2 \cdot \mathcal{GD}_2\left(\mathcal{F}_1,\mathcal{F}_2\right) \tag{11}$$

with $\lambda_1 > 0$ and $\lambda_2 > 0$. The combination of $\mathcal{GD}_1$ and $\mathcal{GD}_2$ is what we call *Sketch Distortion* (SKD). Figure 3 gives a visual understanding of this concept. Figure 3-1) shows two instances of face images from subject A, while faces in figure 3-2) belong to subject B. The visual geometric difference between the two persons is reflected in the Sketch Distortion term, whose values are shown in table 1 ($\lambda_1 = \lambda_2 = 1$).

Table 1. Sketch Distortion (SKD) between the face images from figure 3.

		Subject A		Subject B	
		Image 1	Image 2	Image 1	Image 2
Subject A	Image 1	0	1851	3335	3326
	Image 2	1851	0	3053	2821
Subject B	Image 1	3335	3053	0	1889
	Image 2	3326	2821	1889	0

7 Results

In this section, we present the achieved results using this novel approach. From our previous work ([6]), the matching procedure between points has been improved, and the *Sketch Distortion* term has been introduced. As before, we use the XM2VTS database [7] and the Lausanne protocol (configuration I) [8]. The evaluation set is used to calculate the threshold for which the False Acceptance Ratio (FAR) equals the False Rejection Ratio (FRR) over this set. This threshold (Equal Error Rate threshold) will be employed during the testing phase. The FAR and the FRR over the test set are presented in table 2 for different configurations. In the second row, only textural information is used, i.e. $\lambda_1 = \lambda_2 = 0$ (local textural information is compared at geometrically matched fiducial points).

[3] Note that the centroid of the constellation has been substracted from the point coordinates in order to deal with translation.

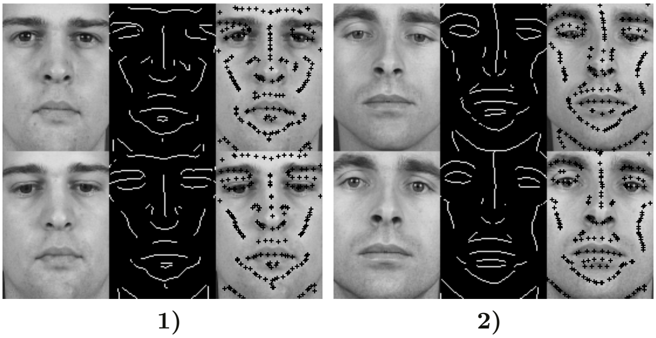


Fig. 3. 1) Top: *Left*: First image from subject A. *Center*: Valleys and ridges sketch. *Right*: Grid adjusted to the sketch. **Bottom**: *Left*: Second image from subject A. *Center*: Valleys and ridges sketch. *Right*: Grid adjusted to the sketch. **2)** Same as **1)** but for subject B.

The results in the third row were achieved by normalizing $\mathcal{GD}_1$ and $\mathcal{GD}_2$ to the range $[0, 1]$. In order to obtain adequate values for λ_1 and λ_2, we performed a grid-search on (λ_1, λ_2) and we chose the pair that minimized the Total Error Rate, i.e. FAR+FRR, over the evaluation set, resulting in $\lambda_1 = \lambda_2 = 0.23$. It is clear that, both the new location and matching of points and the SKD term, properly weighed, help to distinguish between subjects.

Table 2. $FRR_{test}(\%)$ and $FAR_{test}(\%)$ for different configurations.

Method	$FRR_{test}(\%)$	$FAR_{test}(\%)$
Previous work [6]	2.5	8.4
Textural (T)	2.5	6.63
T+SKD	2.5	4.44

8 Conclusions and Further Research

In this paper, we have presented an inherently discriminative approach to automatic face recognition by combining shape and textural information. Fiducial points are located over lines that depict each individual face geometry, and shape differences between constellations of points from two faces are measured using the *Sketch Distortion* term. Gabor jets provide the textural information as defined in [2]. Results over the standard XM2VTS database show that the method is comparable to the best ones reported in the literature and a clear improvement from those reported in [6]. Preliminar experiments have shown that even better results can be achieved by finding the most discriminative fiducial points (using simple Fisher criteria) and performing a linear projection instead of a simple sum of local deformations. Also, we hypothesize that a discriminative combination of the three scores we got, $\mathcal{S}_{\mathcal{J}}$, $\mathcal{GD}_1$ and $\mathcal{GD}_2$, will yield a better performance.

Acknowledgements

We would like to thank Jorge Naranjo-Bouzas for his helpful comments that lead to the originary idea.

References

1. Belongie, S., Malik, J., Puzicha J.: "Shape Matching and Object Recognition Using Shape Contexts," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 24, April 2002
2. Wiskott, L., Fellous, J.M., Kruger, N., von der Malsburg, C.: "Face recognition by Elastic Bunch Graph Matching," IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7), 775-779, 1997
3. Pujol, A., López, A., Alba, José L. and Villanueva, J.J. : "Ridges, Valleys and Hausdorff Based Similarity Measures for Face Description and Matching," Proc. International Workshop on Pattern Recognition and Information Systems, pp. 80-90. Setubal (Portugal), July 2001
4. Pearson, D.E., Hanna, E. and Martinez, K., "Computer-generated cartoons," Images and Understanding, 46-60. Cambridge University Press, 1990
5. López, A. M., Lumbreras, F., Serrat, J., and Villanueva, J. J., "Evaluation of Methods for Ridge and Valley Detection," IEEE Trans. on PAMI, 21(4), 327-335, 1999
6. González-Jiménez, D., Alba-Castro J.L., "Frontal Face Authentication through Creaseness-driven Gabor Jets," in Proceedings ICIAR 2004 (part II), pp. 660-667, Porto (Portugal), September/October 2004.
7. The extended xm2vts database.
   http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/
8. Luttin, J. and Maître, G. "Evaluation protocol for the extended M2VTS database (XM2VTSDB)." Technical report RR-21, IDIAP, 1998.

Compact and Robust Fingerprints Using DCT Coefficients of Key Blocks

Sheng Tang[1,2], Jin Tao Li[1], and Yong Dong Zhang[1]

[1] Institute of Computing Technology, Chinese Academy of Sciences
100080, Beijing, China
{ts,jtli,zhyd}@ict.ac.cn
[2] Graduate School of the Chinese Academy of Sciences
100039, Beijing, China

Abstract. In this paper, we present a novel fingerprinting method for image authentication, the fingerprint length of which is very short (only 81 bytes) and independent on image sizes. First, we extract features based on the block DCT coefficients of images, and binarize the feature map to get key blocks. Then we apply Principal Components Analysis (PCA) to the DCT Coefficients of key blocks. Finally, we take the quantized eigenvector matrix (9×9) as fingerprints. Experimental results show that the proposed method is discriminative, robust against compression, and sensitive to malicious modifications.

1 Introduction

Fingerprints are perceptual features or short summaries of a multimedia object [1]. They can be used for identifying contents just as human fingerprints are used for identification. The aim of fingerprinting (also known as multimedia identification, robust hashes, robust signatures, or passive watermarking) is to provide fast and reliable methods for content identification [1]. It is an emerging research area that is receiving increased attention.

A number of applications of multimedia fingerprinting such as multimedia authentication, indexation of content, and management of large database, were detailed in [1–3]. A typical example of application is multimedia authentication, the key issue of which is to protect the content itself instead of the particular representation of the content without access to the original signals [2, 4, 7]. This renders traditional cryptographic schemes using bit-sensitive hash functions not applicable [1, 2, 5], for multimedia signals can be represented equivalently in different forms, and undergo various manipulations during distribution that may carry the same perceptual information. Therefore, fingerprints should be both discriminative and robust [1].

Researchers have paid great efforts on fingerprinting techniques. Up to now, many image fingerprinting methods have been proposed [1–8]. Schneider and Chang [4] proposed a scheme based on image-block histograms to authenticate the content of an image. Although their scheme is compression tolerant, it has two main drawbacks: considerably large storage requirement of histograms and

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 521–528, 2005.

its security due to the easy way to modify an image without changing its histogram. Bhattacha and Kutter [5] proposed a method based on the locations of salient feature points by using a scale interaction model and Mexican-Hat wavelets. Although the extracted fingerprints are very short, the selection process, relevance of the selected points and its robustness to lossy compression are unclear [8]. Moreover, the feature points are too few and separate to capture the major content characteristics from a human perspective. Thus the method is not discriminative and may be inadequate for detecting some modifications inside the objects. Lou DC and Liu JL [6] proposed a method based on quantization and compression of means of all blocks. Because blocks can be easily modified without changing their means, security problem similar to that of [4] still exists. Queluz [7] proposed techniques to generate fingerprints based on moments and edges. Because moments ignore the spatial distribution of pixels, different images may have same or similar moments. Consequently, moment features are not discriminative enough. Additionally, it is easy to modify an image without changing its moments. Several issues have to be further solved such as the reduction of fingerprint length, the consistency of edge detector, and the robustness to color manipulations [8]. Ching-Yung Lin and Shih-Fu Chang [3, 8] present an effective technique for image authentication which can prevent malicious manipulations but allow JPEG lossy compression. However, their extracted fingerprints are not very compact compared with our method, and largely depend on the image sizes and the number of DCT coefficients compared in each block pair. Recently, a compact and robust fingerprinting method based on radon transform has been proposed in [1]. But the method is not intended for authentication, and is not based on the DCT domain. Hence it can not directly extended to compressed video stream.

In this paper, we present a novel fingerprinting scheme based on the DC and low-frequency AC terms of block DCT coefficients. The procedure for generating fingerprints is shown in Fig.1. First, we extract features based on block DCT coefficients, and binarize the feature map to get key blocks. Then we apply PCA to the DCT Coefficients of key blocks (data matrix A in Fig.1). Finally, we take the eigenvector matrix as fingerprints. Experiments show that the proposed method is discriminative, robust against compression, and sensitive to malicious modifications.

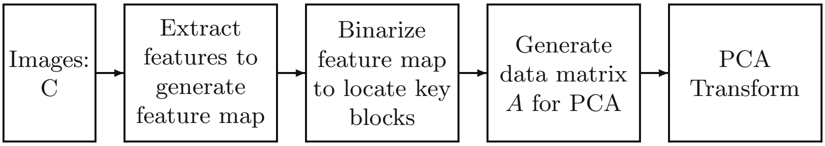

Images: C
Extract features to generate feature map
Binarize feature map to locate key blocks
Generate data matrix A for PCA
PCA Transform

Fig. 1. Procedure for generating fingerprints

As we quantize each element of the eigenvector matrix (9 × 9) to an one-byte integer, the length of extracted fingerprint is very short (only 81 bytes) and independent on image sizes. Since fingerprint lengths of most existing methods depend on image sizes, independence on image size is of great importance, especially for large images, in watermarking-based authentication because the amount of information that can be embedded within an image is limited [3, 7]. Furthermore, since the middle-frequency terms of block DCT coefficients are not used by the proposed method, embedding fingerprint there may be feasible. Additionally, the security problem similar to that of [4] no longer exists, for it is difficult to modify an image without altering its eigenvectors after the zigzag order indices of blocks are taken into consideration in our method.

The remainder of this paper is organized as follows. Section 2 details key block location. Section 3 describes fingerprint generation. Section 4 proposes fingerprint matching method. Section 5 detailes experimental results. Finally, Section 6 summarizes the contributions of this paper and discusses future work.

2 Key Block Location

Hotelling's T-square (HTS) statistic is a measure of the multivariate distance of each observation from the center of the data set, and an analytical way to find the most extreme points in the data. We use the HTS of block DCT coefficients as features of images to locate key blocks.

$$B = \begin{pmatrix} D_{11}\,,\ D_{12}\,, \cdots,\ D_{18} \\ D_{21}\,,\ D_{22}\,, \cdots,\ D_{28} \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ D_{N1}, D_{N2}, \cdots, D_{N8} \end{pmatrix} \tag{1}$$

HTS can be calculated via PCA algorithm. The function princomp() in Matlab Statistics Toolbox describes the PCA algorithm adopted by us, and the fourth output of the function is HTS. We implement it with MATLAB C++ Math Library 2.0. To achieve high speed, instead of using the covariance matrix of B as shown in (1), we adopt the centered and standardized matrix $std(B)$ before using PCA [9]. This is the same with the PCA algorithm in section 3.

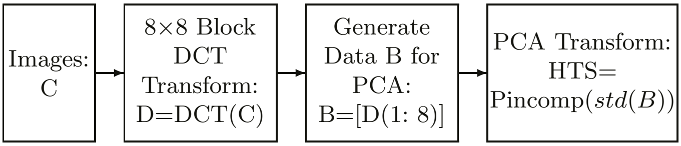

Images: C
8×8 Block DCT Transform: D=DCT(C)
Generate Data B for PCA: B=[D(1: 8)]
PCA Transform: HTS= Pincomp($std(B)$)

Fig. 2. Procedure for computing HTS

To calculate HTS, we apply PCA to the DC and low-frequency terms of block DCT coefficients. The procedure for calculating HTS is shown in Fig.2. First, we transform the original image into 8×8 block-DCT domain. Then, we prepare data matrix B for PCA: divide the DC and 7 low-frequency AC terms (as shown in Fig.3) by the corresponding values of the quantization table in the

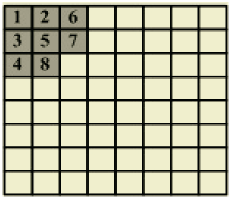


Fig. 3. 8×8 block-DCT coefficients used to compute HTS. The upper-left element of this figure corresponds to the DC term of each 8×8 block DCT. The shaded elements indicate the set of coefficients used in our method, and the numbers indicate the column indices in the data matrix for PCA

JPEG compression process, and place them into a N×8 matrix in zigzag order, where N is the total number of blocks. This can be represented as (1), where D_{ij} denotes the j^{th} DCT coefficients of the i^{th} block of the image C.

The HTS values of all blocks compose the feature map. After the HTS feature map is formed, we binarize it to get key blocks. Before binarization, we quantize each element HTS_i of HTS to an integer $HTS_q \in [0, 127]$ according to (2):

$$HTS_q = \lfloor \frac{127(HTS_i - HTS_{min})}{HTS_{max} - HTS_{min}} \rfloor \tag{2}$$

where HTS_{max}, HTS_{min} mean the maximum and minimum values of HTS.

After quantization, all the HTS values can be classified into two categories C_1: $\{0, k\}$ and C_2: $\{k+1, 127\}$, where k is the threshold of binarization. Thus, we can define the between-class variance as (3).

$$\sigma_b^2 = [u_1(k) - u_2(k)]^2 (\sum_{i=0}^{k} P_i)(\sum_{i=k+1}^{127} P_i) \tag{3}$$

where, P_i is the probability defined as (4), n_i is the number of blocks whose quantized HTS_q equal $i (i = 0, \ldots, 127)$,

$$P_i = \frac{n_i}{N} \tag{4}$$

and u_1, u_2 are means of C_1 and C_2 defined as (5) and (6).

$$u_1(k) = \frac{\sum_{i=0}^{k} iP_i}{\sum_{i=0}^{k} P_i} \tag{5}$$

$$u_2(k) = \frac{\sum_{i=k+1}^{127} iP_i}{\sum_{i=k+1}^{127} P_i} \tag{6}$$

Consequently, we can determine the HTS threshold k_b for binarization by (7), and take the blocks whose HTS_q are greater than k_b as the key blocks.

$$k_b = argmax\{\sigma_b^2\} \tag{7}$$

3 Fingerprint Generation

As shown in Fig.1, after key block location, we prepare the data matrix A for PCA as shown in (8), where D_{ij} denotes the j^{th} DCT coefficients of the i^{th} key block of the image, and K is the number of all key blocks, Z_i is the zigzag order index of the i^{th} key block for consideration of its spatial location.

$$A = \begin{pmatrix} D_{11}\,, \cdots, D_{18}\,, Z_1 \\ D_{21}\,, \cdots, D_{28}\,, Z_2 \\ \dots\dots\dots\dots\dots\dots \\ D_{K1}, \cdots, D_{K8}, Z_K \end{pmatrix} \tag{8}$$

Then, we apply PCA to the matrix A, and use the resultant 9×9 eigenvector matrix V as fingerprint. For more compact fingerprint, compression techniques such as quantization and entropy coding, and Discrete Fourier Transform, can be used to reduce the length of the fingerprint if necessary. In our experiments, we quantize each element a of fingerprint V to an one-byte integer a_q by:

$$a_q = \lfloor 127(1 + a) \rfloor \tag{9}$$

For image authentication, a secure fingerprint can be achieved by using a private key to perturb all the elements of the quantized fingerprint matrix in a random order, and encrypt the fingerprint for secure communication.

4 Fingerprint Matching

Two images are declared similar (for indexation) or authentic (for authentication) if the similarity S between their fingerprints is above a certain threshold T. The main idea of fingerprint matching is that if two images are considered similar or authentic, corresponding eigenvectors from the two fingerprints should be high correlative. Thus, S can be calculated by computing correlation between each pair of eigenvectors, that is, the cosine of the angle between them.

After dequantizing each element a_q of eigenvector matrices by:

$$a = \frac{a_q}{127} - 1 \tag{10}$$

we let $V_o = (\alpha_{o1}, \alpha_{o2}, \dots, \alpha_{o9})$ and $V_t = (\alpha_{t1}, \alpha_{t2}, \dots, \alpha_{t9})$ be the dequantized eigenvector matrices of the original image C_o and the image C_t to be tested respectively. Since eigenvector matrix is orthogonal, S can be mathematically calculated by computing the arithmetic mean of correlations of all the pairs as:

$$S = \frac{1}{9} \sum_{i=1}^{9} |\alpha'_{oi} \alpha_{ti}| \tag{11}$$

where α'_{oi} denotes the transpose of column vector α_{oi}.

5 Experimental Results

In evaluating our proposed method, we tested it on the well-known "F14" image (732 × 500) and the 2000 test images randomly selected from the Corel Gallery database including many kinds of images (256 × 384 or 384 × 256). Prior to extracting fingerprints, we normalized all the images by taking the luminance component although it can be applied to other components. Resizing is not necessary because fingerprint lengthes are independent on image sizes. To do experiments, we first extracted fingerprints from the 2000 images.

To test robustness of the method, we used StirMark [10] to compress the 2000 images to various JPEG images with different quality levels Q ranging from 10% to 90%, and calculated S between images and their corresponding JPEG images. The mean and standard deviation (Std) of the measured S were shown in Table.1. It shows that our method is fairly robust against compression.

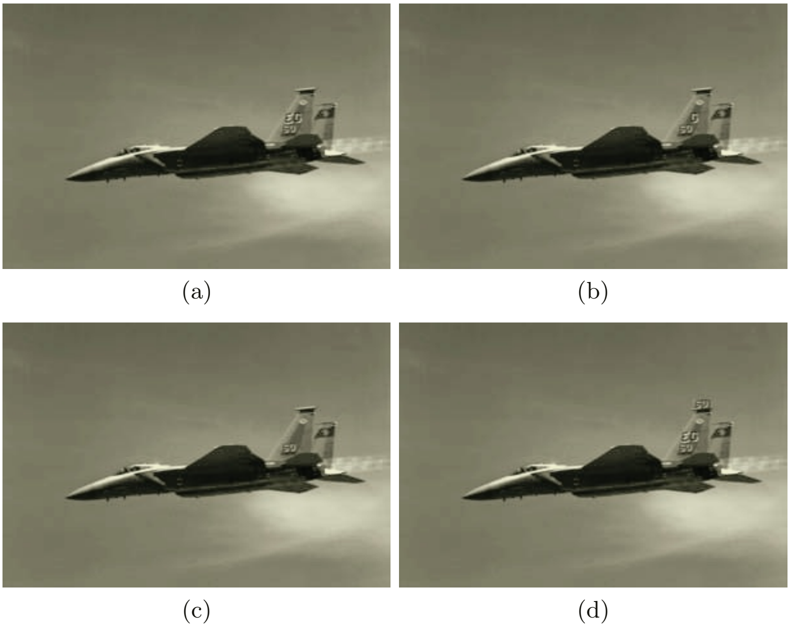
Table 1. Mean and Std of the measured S between images and corresponding JPEG images for 2000 test images

JPEG Compression	Mean	Std
JPEG(Q=10%)	0.8361	0.1367
JPEG(Q=20%)	0.9167	0.0970
JPEG(Q=30%)	0.9411	0.0821
JPEG(Q=40%)	0.9554	0.0688
JPEG(Q=50%)	0.9631	0.0635
JPEG(Q=60%)	0.9664	0.0533
JPEG(Q=70%)	0.9751	0.0467
JPEG(Q=80%)	0.9856	0.0375
JPEG(Q=90%)	0.9887	0.0306

For image authentication, since obvious degradation exists in many images of JPEG(Q=10%), we can set the mean S (0.9167) of JPEG(Q=20%) as the threshold T, or even greater according to various applications.

We made small modifications of "F14" as shown in Fig.4. The measured S between the original image and the tampered images (b), (c) and (d) are 0.8409, 0.7830 and 0.8077 respectively. All those values are below the threshold $T = 0.9167$. Thus, we successfully detected that the three images were tampered. It shows that our method is sensitive to malicious modifications of images.

To test discriminability of the method, we randomly selected 262140 pairs of fingerprints of the 2000 test images, and calculated S between each pair. The histogram of the measured S was shown in Fig.5. All the measured S were in the range between 0.0710 and 0.8440. The mean and standard deviation were 0.3723, 0.1005. We can see that the histogram closely approaches the ideal random i.i.d. case N(0.3723, 0.1005). The mean of the measured S was far below $T = 0.9167$. Thus we can arrive at very low false alarm rate (the probability that declare two different images as similar or authentic): 3.0318×10^{-8}. The above results show that our method is discriminative.



(a) (b)

(c) (d)

Fig. 4. Authentication test on the image “F14”: (a) Original image “F14”; (b) Remove the char “E” on the empennage; (c) Remove the two chars “EC” on the empennage; (d) Copy the two chars “GO” onto the top of the empennage

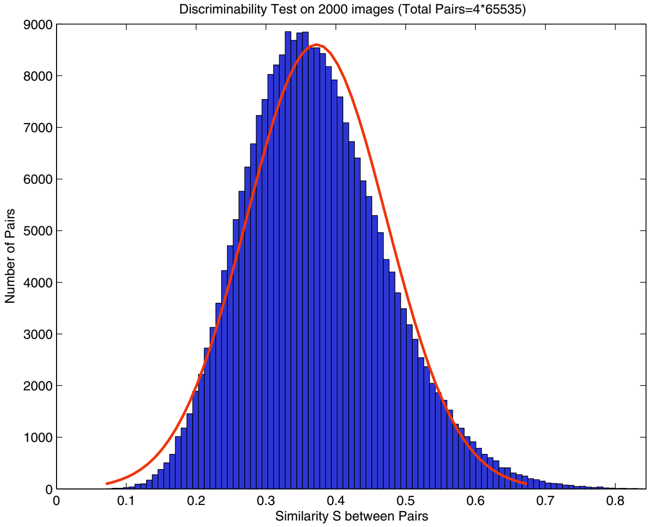


Fig. 5. Histogram of the measured S between 262140 pairs of images randomly selected from 2000 images. The red line represents the ideal random i.i.d. case N(0.3723, 0.1005)

6 Summary and Conclusions

In this paper, we present a novel image fingerprinting method for authentication based on PCA of the DCT coefficients of key blocks determined by the HTS threshold. Experiments show that the proposed method is discriminative, robust against compression, and sensitive to malicious modifications. It is convenient to extend our method to verify compressed video streams without DCT transforms. Furthermore, since the fingerprint length is only 81 bytes long regardless of image sizes, and the middle-frequency terms of block DCT coefficients are not adopted by our method, combining our method with semi-fragile watermarking and embedding fingerprint there may be feasible.

Acknowledgement

This work is supported by National Nature Science Foundation of China under grant number 60302028.

References

1. J.S.Seoa, J.Haitsmab, T.Kalkerb, C.D.Yoo, "A robust image fingerprinting system using the Radon transform," *Signal Processing: Image Communication*, 19(4):325-339, April 2004.
2. B.B.Zhu, M.D.Swanson, A.H.Tewfik, "When seeing isn't believing - multimedia authentication technologies", *Signal Processing Magazine, IEEE,* 21(2):40-49,2004.
3. C.-Y. Lin and S.-F. Chang, "Robust digital signature for multimedia authentication", *Circuits and Systems Magazine, IEEE,* 3(4):23-26, 2003.
4. M. Schneider, S.-F. Chang, "A robust content based digital signature for image authentication", *In: Proceedings of IEEE ICIP 96, Lausanne, Switzerland,* Vol.3:227-230, October 1996.
5. S. Bhattacharjee, M. Kutter, "Compression tolerant image authentication", *In: Proceedings of the IEEE ICIP 1998, Chicago, IL,* October 1998.
6. Lou DC, Liu JL. "Fault resilient and compression tolerant digital signature for image authentication", *IEEE Transactions on Consumer Electronics,* 46(1):31-39, 2000.
7. M.P. Queluz, "Authentication of digital images and video: generic models and a new contribution", *Signal Processing: Image Communication,* 16(5):461-475, Jan. 2001.
8. C.-Y. Lin and S.-F. Chang, "A robust image authentication method distinguishing JPEG compression from malicious manipulation", *IEEE Trans. Circuits Syst. Video Technol.*, 11(2): 153-168, 2001.
9. J. Edward Jackson, *A User's Guide to Principal Components,* John Wiley & Sons, Inc., pp. 1-25, 1991.
10. Fabien A. P. Petitcolas, "Watermarking schemes evaluation", *IEEE Signal Processing,* 17(5):58-64, September 2000.

Fingerprint Matching Using Minutiae Coordinate Systems

Farid Benhammadi, Hamid Hentous, Kadda Bey-Beghdad, and Mohamed Aissani

Military. Polytechnic School, Laboratory of Computer Science
BP 17, Bordj-El-Bahri 16111 Algiers, Algeria
{benhammadif,hentoush,beghdadbey,maissanim}@yahoo.fr

Abstract. In this paper, we propose an original fingerprint matching algorithm using a set of intrinsic coordinate systems which each one is attached to each minutia according to its orientation estimated from fingerprint image. Exploiting these coordinate systems, minutiae locations can be redefined by means of projection of these minutiae coordinates on the relative reference of each orientation minutia. Thus, our matching algorithm use these relative minutiae coordinate to calculate the matching score between two fingerprints. To avoid the directional field estimation errors, we propose a minutia orientation variation in order to manage the projection errors of the location minutiae. With this technique, our approach doesn't embedding fingerprint alignment into the minutia matching stage to design the robust algorithms for fingerprint recognition. The algorithm matching was tested on a fingerprint database DB2 used in FVC2000 and the results are promising.

1 Introduction

A number of biometric technologies have been developed and several of them are being used in a variety of applications. Among these, fingerprint is one that is most popular and commonly used. So, the fingerprint recognition is a complex pattern recognition problem, designing algorithms of extracting fingerprint features and matching them for automatic personal identification. The existing popular fingerprint matching approaches can be classified into three families: correlation-based matching [1] and ridge feature-based matching [2] and minutiae-based matching [3,4,5]. This last is regarded as the most popular and widely used approaches.

Majority of these approaches do not avoid pre-alignment stage to recover the geometric transformation between the two fingerprint impressions in order to ensure an overlap of common region before the minutiae pairing. This fingerprint pre-alignment is certainly primordial in order to maximize the number of matching minutiae. These geometric transformation parameters are generally using a Generalized Hough Transform [4].

The disadvantage of the preceding approaches is the fact that they have not avoided the pre-alignment stage to estimate the geometric transformation parameters which are computationally expensive. To overcome this problem, some authors perform minutiae matching locally [6,7] for an avoiding pre-alignment stage. A few other attempts have been proposed that try to globally match minutiae without making the recourse to fingerprint pre-alignment [8]. These last authors introduce an intrinsic

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 529–536, 2005.

coordinate system based on portioned regular region defined by the orientation field and the minutiae are defined with respect to their position in this coordinate system. So, this approach has some practical problems such as reliably partitioning the fingerprint in regular regions and unambiguously defining intrinsic coordinate axes in poor quality fingerprint images [9].

This study described an original fingerprint matching approach using the local coordinate systems for fingerprint minutiae. Rather than aligning the areas containing minutiae, a local coordinate system aligned with the dominant ridge orientation is proposed to be stored with each minutiae entry. Expressing other minutiae in this local system yields a translation and rotation-invariant feature location.

The rest of this paper is organized as follows. First, Section 2 introduces the orientation-based minutia coordinate system, which is used in Section 3 for the minutiae matching algorithm. Finally, Section 4 presents the experimental results and discussions.

2 An Orientation-Based Minutiae Coordinate System

The large number of approaches to fingerprint features extraction can be coarsely categorized into two families. The first one transforms the gray level fingerprint images into the thinned binary images to which the features are then extracted. However the binary images may lose a lot of feature information, and are sensitive to noise. The second one exploits the direct gray-scale image extraction [9]. So, our feature extraction method is a light modification of this last method using our ridge orientation definition.

In our approach, we use the two minutiae feature parameters:

- x and y coordinate of the minutia point
- θ the minutia orientation which is defined as the local associated ridge-valley orientation.

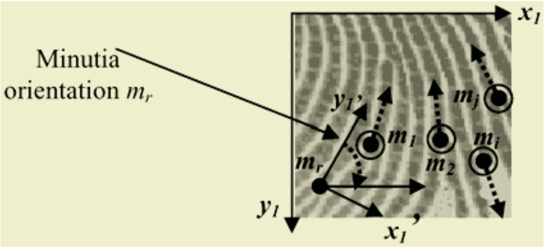


Fig. 1. The OMCS for the minutia m_r

The first step in our method, we start by modifying the orientation of all detected minutiae. By definition, the classical[1] ridge-valley orientation values are included in the interval $[-\pi/2, \pi/2]$. In our method, we change the sense of the classical orientation according to the ends abruptly and the convergences bifurcations. So, two

[1] We use the averaging of square gradient to estimate the classical orientation [8].

ridges of opposite directions θ and θ+π , respectively are not both along a line of orientation θ. Thus, in our approach, the ridge-valley orientation values are included in the interval $[-\pi\ ,\ \ \pi\ ]$.

The second step is to construct the orientation-based minutia coordinate system reference (OMCS). This coordinate system is defined in each minutia point by two reference axes, which is illustrated in Fig 1. Thus, the first reference axis is given by the fixed sense orientation θ, while the second reference axis is given by the perpendicular line (clockwise sense). The origin of the OMCS is minutia point.

According to our ridge orientation definition, all relative minutiae features in the each OMCS remain unchanged for translation and rotation of the fingerprint. This is illustrating by Fig 2. which a rotation is applied for the fingerprint image.

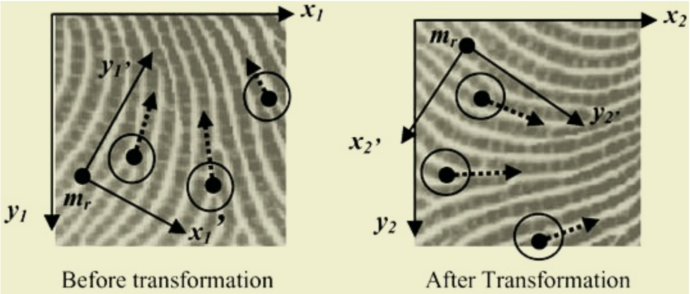


Fig. 2. The unchanged relative minutiae features according the OMCS

The last step consists in projecting the initial minutiae localizations in these new OMCS. This geometric transformation can be estimated as follows. Consider a fingerprint that is represented by their set of minutiae $T = \{\ m_i\ i = 1 \cdots n\ \}$ where n is the number of the minutiae and let m_r is a reference minutia. Let θ_{m_r} represent the original m_r orientation according the initial coordinate system. Thus, the transformation of the minutia mi such that $i \neq r$ is defined by:

$$\begin{bmatrix} x' \\ y' \\ \theta' \end{bmatrix} = \begin{bmatrix} \cos(\theta_{m_r}) & -\sin(\theta_{m_r}) & 0 \\ \sin(\theta_{m_r}) & \cos(\theta_{m_r}) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta\theta \end{bmatrix} \quad \text{where} \quad \begin{cases} \Delta x = x_{m_r} - x_{m_i} \\ \Delta y = y_{m_r} - y_{m_i} \\ \Delta\theta = \theta_{m_r} - \theta_{m_i} \end{cases} \qquad (0)$$

As result, we obtain a minutia m_r projection feature vector (noted P_{m_r}) that describes its global characteristics. Its contains the relative orientation θ_i^r and polar coordinates ρ_i^r and φ_i^r of each minutia m_i for $i \neq r$ according to OMCS of the reference minutia m_r and ridge counter R_c that describes number of ridges between the minutiae m_r and m_i. These features are invariant with respect to OMCS alignment if we have a robust minutia extraction module.

3 Feature Matching in Fingerprint Images

In Section 2, we describe a method of minutiae matching using the proposed minutia relative features computed from each OMCS. In the rest of the paper, all projection feature vectors (Cartesian coordinates) will be converted to the polar coordinates. Let $T = \{ m_i \; i = 1 \cdots n \}$ and $I = \{ m_j^{'} \; j = 1 \cdots m \}$ be the minutiae lists extracted from the template and input fingerprint, respectively. For any pair minutiae m_i and $m_j^{'}$, the similarity can been established based on their projection feature vectors $P_{m_j^{'}}$ and P_{m_i} if there are at least two elements which are similar according the following constraints on the polar coordinates and orientation difference:

$$\left| \rho_i^r - \rho_j^{r'} \right| \leq T_\rho \tag{1}$$

$$\left| \varphi_i^r - \varphi_j^{r'} \right| \leq T_\varphi \tag{2}$$

$$\left| \theta_i^r - \theta_j^{r'} \right| \leq T_\theta \tag{3}$$

where T_ρ, T_φ and T_θ are the acceptable tolerances of parameters errors polar coordinates and orientation difference respectively.

Table 1. The projection errors according the orientation precision

Orientation precision (by degrees)	Radius ρ_i^r between the minutiae m_i and $m_j^{'}$ (by pixels)					
	12	**25**	**50**	**100**	**150**	**200**
1°	0,209	0,436	0,872	1,745	2,618	3,491
5°	1,046	2,181	4,362	**8,724**	**13,09**	**17,45**

Intuitively, we allow some degree of tolerance in matching the minutiae according to these tolerances. But this does not manage the projection errors when the two minutiae are far apart in each OMCS. The Table 1 show the average projections errors (the radius ρ) obtained according to the orientation precision. For example, consider T_ρ= *0.5* as the tolerance error. This value does not allow pairing minutiae for the range [-5°, 5°] as orientation errors for all locations having the distance≥100 pixels. This shows that these locations are not invariant with respect to OMCS references even we have a robust minutia extraction module.

In order to define an effective matching approach, we propose a consolidation step to tolerate local distortion and minutiae extraction errors by the orientation field estimation and projections. This will be defined in what follows. We propose the orientation variation to correct these features errors. This correction is defined by a precision orientation variation of $\pm \Delta\theta$ degrees as shown in Fig3. This variation precise the possible orientation angles of the minutiae. Then for each discredited orientation value θi in the interval [θ-Δθ, θ+Δθ], we redefined a novel OMCS compared to this new orientation and we recomputed the relative minutiae coordinates again.

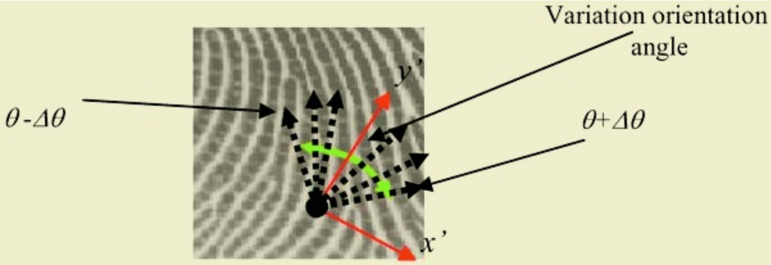


Fig. 3. The minutia orientation variation

Table 2. The projection vector for minutia 4 of template and input fingerprints

Minutia number	Template fingerprint minutiae projections				Input fingerprint minutiae projections			
	Polar coordinates				Polar coordinates			
	ρ_i^4	φ_i^4	θ_i^4	Rc	ρ_j^4	φ_j^4	θ_j^4	R_C
1	24,698	1,198	-0,3793	2	25,080	1,491	-0,3793	2
2	52,154	-0,567	-0,6283	4	53,000	-0,557	-0,6283	4
3	100,125	0,050	-0,7732	9	99,126	0,050	-0,7732	9
4 (m_r)	**0,000**	**0,000**	**0**	**0**	**0,000**	**0,000**	**0.0132**	**0**
5	163,110	0,037	-0,7305	17	173,416	0,069	-0,7305	17
6	216,906	-0,575	2,2439	22	287,312	-0,501	2,2339	29
7	278,548	0,717	-0,6301	29	258,118	0,267	-0,6101	26
8	184,092	-1,163	-0,6363	20	206,422	-1,157	-0,6363	19

To facilitate the comprehension of this technique, we give a small example of matching in order to find the most plausible alignment. Consider the relative minutiae features illustrated in Table 2.

Let us reconsider the previous tolerances T_ρ, T_φ and T_θ. These fingerprints do not have a maximum pairing because the four relative features of the minutiae 5, 6, 7 and 8 do not satisfy the preceding constraints (1, 2 and 3). So, after a light correction of the orientation angle (approximately 2 degrees) of the minutia 4 of input fingerprint, we obtain the values illustrated in the Table 3.

The best matching degree of the template and input fingerprints, called matching score M_S, will then be determined by the sum of all number of similar elements of the two projection feature vectors $P_{m'_j}$ and P_{m_i} according to each OMCS. Let $\iota(m_i, m'_j)$ be the indicator function that returns the number of corresponding elements of the two projections vectors under the previous constraints. Then the matching score M_S is defined by the following formula:

$$M_S = \left(\alpha + \frac{(1-\alpha)}{n}\right)\sum_{j=1}^{m}\sum_{i=1}^{n}\iota\left(m_i, m'_j\right) \tag{4}$$

where α was set to 0.99

Table 3. The projection vector after correction

Minutia number	Input fingerprint minutiae Projections			
	Polar coordinates			
	ρ_j^4	φ_j^4	θ_j^4	R_C
1	24,698	1,198	-0,3793	2
2	51,313	-0,577	-0,6283	4
3	100,125	0,050	-0,7732	9
4	0,000	0,000	0.0132	0
5	164,110	0,037	-0,7305	17
6	216,781	-0,548	2,2339	22
7	278,548	0,717	-0,6101	29
8	184,092	-1,163	-0,6363	20

In this matching score, we give a greater weight to the number of paired minutiae (α) and the coefficient (1-α) allows to decide between two identification rates when one finds the same number of paired minutiae. Thus, we classify the identification results according to the rate of paired minutiae compared to the whole of the reference minutiae of the fingerprint database.

4 Experimental Results

In our experiments, we have used the DB2 from fingerprint databases used in the 2000 Fingerprint Verification Competition [9]. It contains 800 fingerprint images with 100 distinct fingers, each finger having 8 impressions. The matching performances (FNR: False Negative Rate and FPR: False Positive Rate) achieved on DB2 for a several variations of the minutia orientation are shown in Table 4.

Table 4. The verification results and Time matching on DB2

Range error	Without range		range = [-15, 15]		Average match time (s)
Paired minutiae	FNR	FPR	FNR	FPR	
6 minutiae	15.26 %	25 %	5.26 %	10 %	0.063
9 minutiae	35.57 %	0,05 %	14.57 %	0.02 %	0.065

The Fig4. show a correctly matched pair between two fingerprints using our matching method. In this case, 9 minutiae were necessary to validate paired fingerprint.

The effectiveness of our method decreases with a minimal number of minutiae and fails in the identification because this number is insufficient for minutiae pairing of two fingerprints. This situation reflects a false negative rate (false identification) which caused by the minimal number of the paired minutiae. We noted that incorrect matching is generally caused by the minutiae extraction module (Approximately 11% of extraction errors between the false, missing minutiae and change minutiae type). But this problem may be solved exploiting the differential correlation which is computed in a neighborhood of each paired minutiae by using the approach proposed in [10].

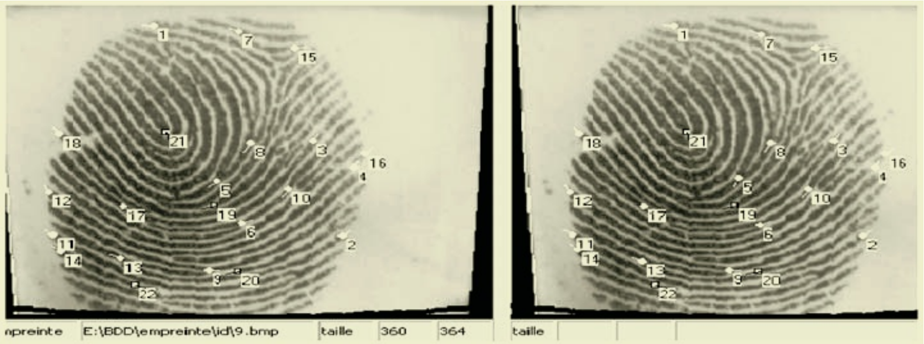


Fig. 4. Example of correctly identified fingerprints

5 Conclusion

This work described an original fingerprint matching technique using an orientation-based minutia coordinate system which is attached to each detected minutia. The main advantage of our approach is the elimination of the fingerprint pre-alignment compared to the precedent approaches. In addition, our approach does not have problem of the definition of an intrinsic coordinate system as the approach described in [8] because we do not use the partitioning the fingerprint in regular regions but we use at least one intrinsic coordinate system among all attached to minutiae. The usefulness of this approach was confirmed in the experiments conducted here, which reveals that the results are encouraging and our approach is promising.

References

1. C. Wilson, C. Watson, and E. Paek, "Effect of Resolution and Image Quality on Combined Optical and Neural Network Fingerprint Matching," Pattern Recognition, vol. 33, no. 2, pp. 317-331, 2000.
2. A. Ceguerra and I. Koprinska, "Integratin Local and Global features in Automatic Fingerprint Verification", In Proc. Int. on Pattern Recognition (16th), vol. 3, no. 2, pp. 347–350, 2002.
3. M. Tico and P. Kuosmanen, "Fingerprint Matching Using an Orientation-Based Minutia Descriptor " IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 25, no. 8, pp. 1009-1014, August 2003.
4. A.K. Jain, L. Hong, S. Pankanti, and R. Bolle, "An Identity-Authentication System Using Fingerprints," Proc. IEEE, vol. 85, no. 9, pp. 1365-1388, 1997.
5. K. Jain, L. Hong, and R. Bolle, "On-Line Fingerprint Verification," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 4, pp. 302-313, Apr. 1997.

6. Zsolt Miklos Kovacs-Vajna,,“ A Fingerprint Verification System Based on Triangular Matching and Dynamic Time Warping,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 11, pp.1266-1276, Nov. 2000.
7. X. Jiang and W.Y. Yau, “Fingerprint Minutiae Matching Based on the Local and Global Structures”, in 15th Proc. Int. Conf. on Pattern Recognition, vol. 2, pp. 1042-1045, 2000.
8. A.M. Bazen and S.H. Gerez. “An Intrinsic Coordinate System for Fingerprint Matching”. In 3rd International Conference on Audio and Video-based Biometric Person Authentifcation, Halmstad, Sweden, June 6-8, 2001.
9. D. Maltoni, D. Maio, A.K. Jain and S. Prabhakar, “Handbook of Fingerprint Recognition”, Sringer Verlag, New York, 2003.
10. T. Hatano, T. Adachi, S. Shigematsu, H. Mirimora, S. Onishi, Y Okasaki and H. Kyuragi, “A Fingerprint Verification Using the Differential Matching Rate,”, Proc. Int. Conf on Pattern Recognition, vol. 3, pp. 799-802, 2002.

The Contribution of External Features to Face Recognition

Àgata Lapedriza, David Masip, and Jordi Vitrià

Computer Vision Center-Dept. Informàtica
Universitat Autònoma de Barcelona
08193 Bellaterra, Barcelona, Spain
{agata,davidm,jordi}@cvc.uab.es

Abstract. In this paper we propose a face recognition algorithm that combines internal and external information of face images. Most of the previous works dealing with face recognition use only internal face features to classify, not considering the information located at head, chin and ears. Here we propose an adaptation of a top-down segmentation algorithm to extract external features from face images, and then we combine this information with internal features using a modification of the non parametric discriminant analysis technique. In the experimental results we show that the contribution of external features to face classification problems is clearly relevant, specially in presence of occlusions.

1 Introduction

During the past several years, face recognition has received significant attention as one of the most successful applications of image analysis and understanding. Several recognition systems have been recently developed due to their usefulness in a lot of real world applications in different areas such as security, entertainment or user friendly interfaces. However, their success is limited by the conditions imposed by many real applications. Recognition of face images acquired in an outdoor environment with changes in illumination, changes in pose or even occlusions is still an unsolved problem and these methods are still nowadays far away from the capability of the human perception system.

Most of the current face recognition algorithms can be categorized into two classes: geometry feature-based and image template based. The geometry feature-based methods analyze explicit local facial features, and their geometric relationships. The template based methods [1] compute the correlation between a face and one or more model templates to estimate the face identity. Statistical tools such as Support Vector Machines (SVM) [2], Linear Discriminant Analysis (LDA) [3], Principal Component Analysis (PCA) [4, 5], Kernel Methods [6] and Neural Networks [7] have been used to construct a suitable set of face templates that can be viewed as features. These kind of methods have proved to be effective in experiments with large databases.

Face recognition applications related to security need to focus on features difficult to imitate and for this reason the recognition systems tend to use only

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 537–544, 2005.

internal face features. Nevertheless, as technology evolves, it is easier to find electronic devices in our everyday life and in this context new applications dealing with face classification have appeared. In these cases the reasons to use only internal features are not longer valid, and given that the contribution of external features in face recognition is notable [8], their use in automaic systems should be revised. In this paper we propose a method to extract external information in face images and an algorithm to combine the information of external and internal features to solve the recognition problem.

The paper is organized as follows: in section 2 we explain the discriminant analysis algorithm to recognize subjects using only internal features. Section 3 shows how we build a model of known external features, how we use this model for the reconstruction of an unseen image and how we classify the unseen image from its reconstruction. Section 4 describes our experiments and section 5 conclude this work.

2 Extraction of Discriminant Internal Features

The extraction of the internal features in our comparative study has been performed using standard linear techniques found in the literature. One of the most used feature extraction algorithms is Principal Component Analysis (PCA) [4], where we obtain the orthogonal set of basis that preserve the maximum amount of data variance. The original $D-$dimensional data can be reconstructed using M coefficients ($M < D$) minimizing the reconstruction error. Nevertheless, sometimes the features that minimize the reconstruction error are not necessarily the features most suitable for classification [11]. If the class labels of the training sample are taken into account, other linear projections can yield a better classification accuracy even though the reconstruction error is not minimized. In this work we have used a modification of the nonparametric discriminant analysis (NDA) algorithm for this purpose. Below, we briefly describe the classic fisher linear discriminant analysis (FLD) [9] technique and the assumptions performed on the training data, to introduce later the NDA algorithm by Fukunaga et al [10], and the modified version used in this work.

We will assume a Nearest Neighbor classifier in this work, given that the feature extraction performed by the NDA algorithm is specially suitable for the NN rule using euclidean distance.

2.1 Discriminant Analysis

Here we look for a transformation matrix $\mathbf{W}$ which maximizes

$$\mathcal{J} = tr((\mathbf{W}\mathbf{S}_W^{-1}\mathbf{W}^T)(\mathbf{W}\mathbf{S}_B\mathbf{W}^T)) \tag{1}$$

Here $\mathbf{S}_B$ and $\mathbf{S}_W$ are the between-class and the within-class scatter matrix respectively. This problem has an analytical solution [11]. $\mathbf{W}$ is constructed using as its rows the M eigenvectors corresponding to the largest M eigenvalues of $\mathbf{S}_W^{-1}\mathbf{S}_B$.

This approach for calculating the within- and between-class scatter matrices makes use of only up to second order statistics of the data. This was proposed in the classic paper by Fisher [9] and the technique is referred to as Fisher Linear Discriminant Analysis (FLD). In FLD the within class scatter matrix is computed as the weighted sum of the class-conditional sample covariance matrices. If equal priors are assumed for the classes C_k, $k = 1, \ldots, K$, then

$$\mathbf{S}_W = \frac{1}{K} \sum_{k=1}^{K} \mathbf{S}_k \tag{2}$$

where $\mathbf{S}_k$ is the class-conditional covariance matrix for C_k, estimated from the data. The between class-scatter matrix is defined as,

$$\mathbf{S}_B = \frac{1}{K} \sum_{k=1}^{K} (\mathbf{m}_k - \mathbf{m}_0)(\mathbf{m}_k - \mathbf{m}_0)^T \tag{3}$$

where $\mathbf{m}_k$ is the class-conditional sample mean and $\mathbf{m}_0$ is the unconditional (global) sample mean.

The following two limitations of FLD have to be noted: the rank of $\mathbf{S}_B$ is $K - 1$, so the number of extracted features can be, at most one in a gender recognition problem (with only two classes). Second, the scatter matrices are calculated assuming Gaussian classes. The solution provided by FLD is blind beyond second-order statistics, so this method may be inaccurate for complex classification structures.

2.2 Non-parametric Discriminant Analysis

Fukunaga and Mantock [10] propose a nonparametric discriminant analysis method as an attempt to overcome the two limitations of FLD noted above. In NDA the between-class scatter matrix $\mathbf{S}_B$ is calculated without the assumption of Gaussian classes. This scatter matrix is generally full rank, thus loosening the bound on the extracted feature dimensionality. Below we briefly expose this technique, extensively detailed in [11].

In NDA, the between-class scatter matrix is obtained as an average of N local covariance matrices, one for each point in the data set. This is done as follows. Let $\mathbf{x}$ be a data point in $\mathbf{X}$ with class label C_j. Denote by $x^{\text{different}}$ the subset of the k nearest neighbors of $\mathbf{x}$ among the data points in $\mathbf{X}$ with class labels different from C_j. We calculate the "local" between-class matrix for $\mathbf{x}$ as

$$\varDelta_B^{\mathbf{x}} = \frac{1}{k-1} \sum_{\mathbf{z} \in x^{\text{different}}} (\mathbf{z} - \mathbf{x})(\mathbf{z} - \mathbf{x})^T \tag{4}$$

The estimate of the between-class scatter matrix $\mathbf{S}_B$ is found as the average of the local matrices

$$\mathbf{S}_B = \frac{1}{N} \sum_{\mathbf{z} \in X} \varDelta_B^{\mathbf{z}} \tag{5}$$

We use $k = 1$ in this study, hence $x^{\text{different}}$ contains only one element, $\mathbf{z}_{\mathbf{x}}^{\text{different}}$, and

$$\mathbf{S}_B = \frac{1}{N} \sum_{\mathbf{x} \in X} (\mathbf{x} - \mathbf{z}_{\mathbf{x}}^{\text{different}})(\mathbf{x} - \mathbf{z}_{\mathbf{x}}^{\text{different}})^T. \tag{6}$$

The M eigenvectors corresponding to the largest M eigenvalues of $\mathbf{S}_W^{-1}\mathbf{S}_B$ define the projection matrix $\mathbf{W}$. In [12] they introduced also a non parametric form of the within-class scatter matrix $\mathbf{S}_W$, to extract features more suitable for the nearest neighbor classification. They propose to use

$$\mathbf{S}_W = \frac{1}{N} \sum_{\mathbf{z} \in X} \Delta_W^{\mathbf{z}} \tag{7}$$

where $\Delta_W^{\mathbf{x}}$ is calculated from the set of k nearest neighbors of $\mathbf{x}$ from the same class label, C_j, x^{same}

$$\Delta_W^{\mathbf{x}} = \frac{1}{k-1} \sum_{\mathbf{z} \in x^{\text{same}}} (\mathbf{z} - \mathbf{x})(\mathbf{z} - \mathbf{x})^T \tag{8}$$

For $k = 1$,

$$\mathbf{S}_W = \frac{1}{N} \sum_{\mathbf{x} \in X} (\mathbf{x} - \mathbf{z}_{\mathbf{x}}^{\text{same}})(\mathbf{x} - \mathbf{z}_{\mathbf{x}}^{\text{same}})^T. \tag{9}$$

In this paper we use the modified NDA algorithm as was proposed in [12] (using the local approximations of $\mathbf{S}_B$ and $\mathbf{S}_W$).

3 Extraction of the External Features

To extract the external features from the face images we have adapted a Top-Down Segmentation algorithm [13] to build a model based on a selection of parts from the object, faces in our case. In fact, we consider their segmentation as a reconstruction of the original image. Then, given a new unseen image, we find the subset of these parts that best reconstruct the image. The information of the matching between the parts-based model and the unseen image is used to classify the sample in classes. The set of pieces of the learned model are called in this paper *Building Blocks.*

The algorithm can be divided in two parts: learning the model from sample images, and the reconstruction of a new image using the set of *building blocks.* In the first step the optimal set of fragments from the object are learned, and the later step yields the features useful for classification of each new image.

3.1 Learning the Model

In this step the best an optimal set of fragments from the face images is computed. Usually this is the most time consuming part of the algorithm, although it is performed only once, off line, and using a generic face training set.

Given a training set consisting on face images with only the external characteristics to analyze (set C), and non face images acquired in natural environments (set $\overline{C}$), we generate subimages at sizes ranging from 12×12 to 24×24 from the training set. Each subimage will be a candidate fragment F_i for the final model. For each F_i the maximum values of the normalized correlation NC_i between F_i and each image from C and $\overline{C}$ are computed. The model is built storing the K fragments with best probability to describe the elements of the class C and not $\overline{C}$ $p(NC_i > \theta_i \mid C)$. The value of the threshold θ_i is computed taking into account a predefined number of false positive that can be tolerated $p(NC_i > \theta_i \mid \overline{C}) \leq \alpha$. For each fragment we also store a mask where only the pixels of the object are active.

For the construction of the model we guarantee that the set of fragments that we keep is able to reconstruct the external features of a generic face image. So additional restrictions to the relative position of the fragments are imposed. We discard similar pieces from the same relative position on a face, trying to achieve enough diversity in the fragments that compose the model for the external features. To perform it, we separately compute the model for different parts of the face images: forehead part, left side, right side, and chin part. Although there is some overlapping in these parts we obtain enough variety of pieces from each part in the final model.

3.2 Extraction of the External Features from Unseen Images

Suppose now that we have learned a model of external features. Given an unseen image our goal is to select the set of fragments of the building blocks that best reconstruct the image, and use this reconstruction to recognize or classify the subject. In figure 1 we show an example where it seems reasonable the use of the reconstruction for classification, given that the obtained reconstruction is not affected by image artifacts.

In this study we have proceeded as follows: we have computed the normalized correlation between the new image and each of the fragments of the building blocks and we have encoded the new image as a vector with the correlation values. We have considered this vector as a representation of the external features of the subject in the image, and we have used it to classify.

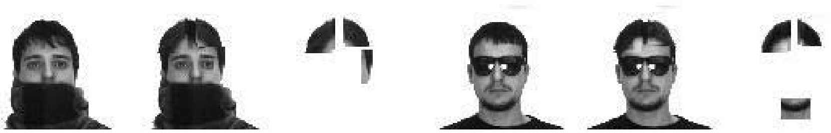


Fig. 1. Example of face images from the AR Face database [14] where different fragments are analyzed at each position. The first image shows the original face, in the second image we plot only the 5 most similar fragments according to the normalized correlation on the face. The third image shows the fragments alone.

3.3 Combination of the Internal and External Information

Once we have obtained the internal and external features for each new unseen image is needed to combine this information in a classification rule. Nevertheless it is not easy to understand the role that the different facial features play in a judgment of identity. In this work we have joined both sets of internal and external features in a single matrix and we have applied standard discriminant analysis techniques such as NDA to select a linear projection that performs the feature extraction from the whole set.

4 Experiments

To show the performance of our purpose in a face recognition problem, we have used the AR Face database [14], which is composed by 26 samples from 126 different subjects. Images were acquired in two different sessions, and for each session there is a frontal face, 3 samples with faces gesturing, 3 samples with illumination changes (frontal and lateral illumination), 3 samples with occlusions produced by the use of glasses (also with frontal and lateral illumination), and 3 samples with occlusions produced by the use of a scarf (with the 3 kinds of illumination). One sample from each type is plotted in table 1 above the results.

We have set the configuration parameters of the experiment as done in [15]. The image set has been split in non overlapping training and test sets, randomly selecting half of the subjects as a generic training data. The testing has been performed on the remaining subjects. A previous preprocess has been preformed on the face database, consisting on resizing each face according to the inter eye distance, and aligning the central position of both eyes. Also the mean has been subtracted from each image. The internal features have been extracted using different linear feature extraction algorithms: PCA and NDA. In all the cases we have selected the central part (33 × 33 pixels) of each image to be used as a input for each internal feature extractor. In the figure 2 we show some examples of the training set.

The external features have been extracted using the algorithm shown in section 3. We have randomly selected 40 generic faces and 40 images with no faces from natural environments. Using this set we have learned the fragments based model, setting the default threshold α to 0.01. Only the 400 fragments with maximum probability have been preserved. We also guarantee that there are fragments from each part of the face in the final set (frontal, chin zone, and both laterals). In the table 1 we compare the results using just internal features with PCA and NDA, with the use of our purpose with external information. We have used the nearest neighbor classifier on each case, with euclidean distance. The optimal dimensionality reduction on each case as been selected cross validating the training data. We show on table 1 the accuracies on each type of image for each case. As can be seen, the NDA algorithm outperforms the PCA in almost all the cases, given that NDA focuses the feature extraction on finding the most discriminative features while PCA just tries to preserve the maximum amount of data variance.

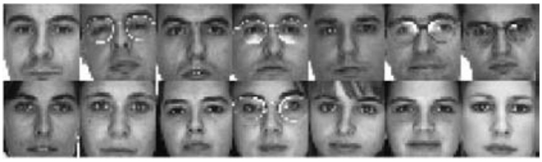


Fig. 2. Central part of some training images with only internal features.

Once we have fixed the NDA technique as the most suitable for the face recognition task, we have combined the external features and the internal ones using this algorithm, in order to obtain higher accuracies combining both approaches (COM in the table 1). The results obtained show that the external features help significantly the face recognition task. Actually, the technique is specially suitable when there are occlusions (sets A08 to A13), as can be seen, the contribution of the external features is notable in this cases, due to the large presence of fragments on the non occluded parts of the face. This allows new features for the NDA algorithm that have not been affected by the occlusions. In frontal and gesture images, the use of external features also improves the NDA using just internal features. In images with strong changes in the illumination, the external features also help when the light is focused on a lateral, giving more importance to the non illuminated side (more fragments). In Images with strong light changes on both sides the external features do not contribute to the global accuracy of the NDA given that the correlation values are mislead in all the cases due to the light.

Table 1. Results using Principal Component Analysis (PCA), Non Parametric Discriminant Analysis (NDA), and our purpose combining internal and external features (COM) for each AR Face image subset (percentage rates).

	 AR01	 AR02	 AR03	 AR04	 AR05	 AR06	 AR07	 AR08	 AR09	 AR10	 AR11	 AR12	 AR13
PCA	76.7	69.7	74.4	51.1	67.4	60.4	60.4	34.9	39.5	34.9	44.2	37.2	23.2
NDA	74.4	79.1	76.7	51.2	76.7	69.8	62.8	34.9	46.5	37.2	62.8	44.2	46.5
COM	81.4	86.0	76.7	55.80	72.1	69.8	74.4	44.2	53.5	41.9	67.4	60.5	51.2

5 Conclusions

In this paper we have proposed a method for face recognition using internal and external features. We have adapted a Parts Based Segmentation algorithm to learn a model to extract the external features of new unseen images. These external features have been combined with the internal ones using classic discriminant analysis techniques. In this work the NDA algorithm has been used for this purpose. The experimental results show that external features improve the ones obtained using only internal information and this indicates that the contribution of external features is relevant for classification purposes, specially when occlusions are present.

As a future work we plan to get better reconstruction methods to represent the external features of each subject using the learned model. Alternative information to the normalized correlation could be used (derivatives, edge detectors). Also the reconstruction could be more reliable if the geometric relationships between the fragments are taken into account. Other algorithms could be used to combine external and internal information, weighting its relative significance.

Acknowledgments

This work has been partially supported by MCYT grant TIC2003-00654 and FP2000-4960 Ministerio de Ciencia y Tecnologia, Spain.

References

1. R. J. Baron. Mechanisms of human facial recognition. International Journal of Man-Machine Studies 15(2):137–178, 1981.
2. V. N. Vapnik. The nature of statistical learning theory. Springer Verlag, Heidelberg, DE, 1995.
3. R. J. Baron. Mechanisms of human facial recognition. International Journal of Man-Machine Studies, 15(2):137–178, 1981.
4. L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. Journal of Optical Society of America, 4(3):519–524, March 1987.
5. M. Turk and A. Pentland. Eigenfaces for recognition. Journal of Cognitive Neuroscience, 3(1):71–86, 1991.
6. Y. Li, S. Gong and H. Liddell. Support vector regression and classification based multi-view face detection and recognition. In IEEE International Conference on Automatic Face and Gesture Recognition, March 2000.
7. A. J. Howell, C. Lee Giles, A. C. Tsoi, and A. D. Back. Face recognition: A convolutional neural network approach. IEEE Transactions on Neural Networks, 8(1):98–113, 1998.
8. I. N. Jarudi and P. Sinha. Relative Contributions of Internal and External Features to Face Recognition. Massachusetts Institute of technology, 2003.
9. R. Fisher.The use of multiple measurements in taxonomic problems. Ann. Eugenics, vol. 7, pp. 179–188, 1936.
10. K. Fukunaga and J. Mantock. Nonparametric discriminant analysis. IEEE Trans. Pattern Anal. Machine Intell., vol. 5, no. 6, pp. 671–678, nov 1983.
11. K. Fukunaga. Introduction to Statistical Pattern Recognition, 2nd ed. Boston, MA: Academic Press, 1990.
12. M. Bressan and J. Vitria. Nonparametric discriminant analysis and nearest neighbor classification. Pattern Recognition Letters, vol. 24, no. 15, pp. 2743–2749, nov 2003.
13. E. Borenstein, S. Ullman. Class-Specific, Top-Down Segmentation. Proceedings of the 7th European Conference on Computer Vision-Part II, 2002, Springer-Verlag.
14. A. Martinez and R. Benavente. The AR face database. Computer Vision Center. Technical Report #24. June 1998.
15. R. Gross, J. Shi and J.Cohn. Quo vadis Face Recognition?, Carnegie Mellon University. Technical report CMU-RI-TR-01-17

Iris Recognition Based on Quadratic Spline Wavelet Multi-scale Decomposition

Xing Ming[1], Xiaodong Zhu[1], and Zhengxuan Wang[2]

[1] Department of Computer Science and Technology, Jilin University
5988 Renmin Street, Changchun 130022, China
mingxing@jlu.edu.cn

[2] Department of Computer Science and Technology, Jilin University
10 Qianwei Street, Changchun 130012, China

Abstract. This paper presents an efficient iris recognition method based on wavelet multi-scale decompositions. A two-dimensional iris image should be transformed into a set of one-dimensional signals initially and then the wavelet coefficients matrix is generated by one-dimensional quadratic spline wavelet multi-scale decompositions. From the basic principles of probability theory, the elements at the same position in different wavelet coefficients matrices can be considered as a high correlated sequence. By applying a predetermined threshold, these wavelet coefficients matrices are finally transformed into a binary vector to represent iris features. The Hamming distance classifier is adopted to perform pattern matching between two feature vectors. Using an available iris database, final experiments show promising results for iris recognition with our proposed approach.

1 Introduction

Biometrics is automated methods of identifying a person or verifying the identity of a person based on a physiological or behavioral characteristic, such as fingerprints, face, iris, handwriting and gait etc. Today the practical applications for biometrics are diverse and expanding, and range from health-care to government, financial services, transportation and public safety and justice.

Since 9.11 terror attack in U.S.A., more reliable and safer security techniques are desired. Due to uniqueness, stability and live status of iris patterns and high reliability and noninvasive acquisition of the iris-based recognition system, iris recognition technology has become a most important biometric solution for personal identification and is becoming an active topic in biometrics [1]–[2].

1.1 Related Work Based on Wavelet

Since the 1990s, much work on iris recognition has been done and great progress has been made [3]–[6]. In recent years wavelet-based technology has been widely used in the field of iris recognition and developed this kind of authentication technology. Boles et al. [5] obtained an iris representation via the zero-crossings of

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 545–552, 2005.

one-dimensional wavelet transform and iris matching was based on two dissimilarity functions. Lim et al. [7] analyzed iris patterns based on Haar wavelet transform and a modified competitive learning neural network (LVQ) was adopted for classification. Ma et al. [8] used wavelet local extremum to locate the local sharp variation points and transformed location information to a binary feature vector. Iris pattern matching was performed by exclusive OR operation.

1.2 Outline

With a one-dimensional (1-D) wavelet multi-scale decomposition, an original signal can be decomposed into one low frequency approximation component and one high frequency detail component at a certain scale. The wavelet coefficients referred to the detail component and those at different scales represent different frequency information of the original signal. Therefore, wavelet coefficients are usually selected to replace the original signal for recognition tasks. The method presented in this paper arranges all the computed wavelet coefficients at different scales into a matrix. Based on the probability theory and lemma, a threshold-based scheme is used to obtain a binary vector to represent iris features. A fast matching scheme based on Hamming distance is adopted to compute the similarity between two binary vectors. Finally we perform a series of experiments to evaluate the proposed approach and provide analysis on the overall experimental results.

The contents of this paper are as follows. Section 2 provides an overview of an iris recognition system working flow. Detailed descriptions of image preprocessing, feature extraction and matching are given in section 3. Experimental results and analysis are stated in Section 4, and finally the conclusions are given in Section 5.

2 How Iris Recognition System Works

Fig. 1 describes the working flow of an iris recognition system. An iris template database being provided in advance, it contains 5 steps, (1) live-acquire an iris; (2) preprocess the iris image; (3) extract iris features and generate a feature vector; (4) match the unknown feature vector against stored templates; (5) provide a matching score to identify an authentication or imposter.

3 Iris Image Analysis and Pattern Matching

3.1 Iris Image Preprocessing

An iris image preprocessing composes of iris localization, normalization and texture enhancement [8]. Iris location can extract the valid iris texture from an iris image including some irrelevant parts (e.g. eyelid, pupil etc). Then the localized iris is unwrapped to a rectangular block of a fixed size in order to reduce the

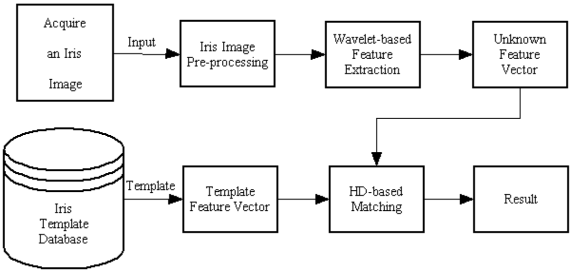


Fig. 1. How an iris recognition system works.

deformation caused by variations of the pupil and obtain approximate scale invariance. Finally, lighting correction and contrast enhancement are applied to compensate for difference of imaging conditions. The whole procedure is illustrated in Fig. 2. From the preprocessing result of an iris image (Fig. 2c) we can see that iris texture characteristics become clearer than those in Fig. 2b.

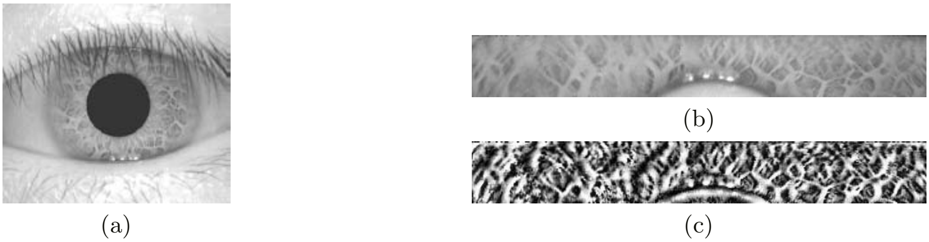


Fig. 2. Iris image preprocessing: (a) original image; (b) normalized image; and (c) enhanced image.

3.2 Iris Feature Extraction

1-D Wavelet Multi-scale Decomposition. A finite 1-D discrete signal s can be expressed as a finite sequence $f_0 = (f_{0,l})(l = 0, 1, \cdots, L-1)$ of a certain length L. With 1-D wavelet multi-scale decompositions, f_0 can be decomposed into M high frequency components $d_1, d_2, \cdots, d_M$ and one low frequency component f_M by the recursions $f_{m+1,l} = \sum\limits_{k=2l}^{2l+N} h_{k-2l} f_{m,k}$ and $d_{m+1,l} = \sum\limits_{k=2l-N+1}^{2l+1} (-1)^k h_{2l+1-k} f_{m,k}$, where $(h_0, h_1, \cdots, h_N)$ is a real-valued sequence used as the wavelet filter with an even length of $N+1$ and the index m is the scale of the wavelet transform. The computed wavelet coefficients over M scales can be arranged into a matrix with the portion W being filled with zero-valued elements as shown in Fig. 3.

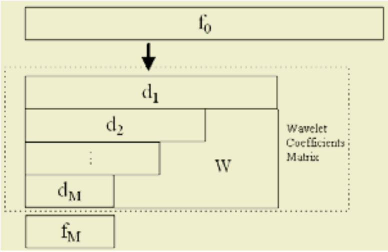


Fig. 3. 1-D wavelet decompositions of a sequence f_0.

Feature Representation. For a representative set of K signals, the computed wavelet coefficients are arranged into a matrix for each individual signal. We denote each such matrix by $B_k=(b_{ij})_k$ $(i=1,2,\cdots,M;j=1,2,\cdots,(L+N-1)/2$ for even L; $j=1,2,\cdots,(L+N)/2$ for odd L; $k=1,2,\cdots,K)$. The elements at the same position in these different matrices from same class can be considered as samples from high correlated variables. Let $\tilde{B}_k=(|b_{ij}|)_k$ and derived from $\tilde{B}_k$ the following standardized matrix G can be represented by:

$$G=(g_{ij})=\frac{\frac{1}{K}\sum_{k=1}^{K}\tilde{B}_k-\mu\left(\frac{1}{K}\sum_{k=1}^{K}\tilde{B}_k\right)\cdot\tilde{I}}{\sigma\left(\frac{1}{K}\sum_{k=1}^{K}\tilde{B}_k\right)} \tag{1}$$

where $\mu(A)$ and $\sigma(A)$ denote the mean and standard deviation of any matrix A, respectively; $\tilde{I}$ is the matrix of the same size as $\tilde{B}_k$ but containing only 1s as its elements. Matrix G is formed by statistical computations of K wavelet coefficients matrices and reflects the whole spectrums of characteristics of K signals.

According to wavelet theory, a wavelet coefficient $d_{(q+1),l}$ with shift l at scale q can be considered as linear combination of many independent random variables. The number of variables is equal to the length of support set of discrete wavelet. In terms of Central Limit Theorem, the distribution of wavelet coefficients matrix satisfies normal distribution. Therefore the elements of matrix G should be should be nearly $N(0,1)$ distributed, Following from the Lemma presented in [9], a threshold of the form $T=\sqrt{2\ln(\tilde{N}/\gamma)}$, with $\gamma\geq e^2$, is applied to the elements of the matrix G, $\tilde{N}$ being the number of computed detailed coefficients. thus we get the binary matrix:

$$G_b=(\Theta(g_{ij}-T)) \tag{2}$$

with the Heavyside function $\Theta(x)=1$ for $x\geq 0$ and $\Theta(x)=0$ for $x<0$. Each element (1 or 0) in G_b represents the correlation degree with the original signal of the corresponding wavelet coefficient and is taken as an iris feature representation.

Feature Vector. As mentioned earlier, a two-dimensional (2-D) iris images should be transformed into 1-D intensity signals initially. Experiments indicate that the informative features in the angular direction corresponding to the horizontal direction in the normalized image have higher intensities than those in other directions [10]. Therefore, a simple horizontal scanning technique was applied to decompose the 2-D normalized image I into a set of 1-D intensity signals S according to the following equations:

$$S_i = \frac{1}{U}\sum_{j=1}^{U} I_{(i-1)*U+j} \quad (i = 1, 2, \cdots, V)$$

$$I = \left(I_1^T, \cdots, I_x^T, \cdots, I_H^T\right)^T \tag{3}$$

where I is the normalized image of $H \times W$, I_x denotes the xth row-data in the image I, U is the total number of rows used to from S_i, V is the total number of 1-D signals.

For K different iris samples from the same class, K series of 1-D intensity signals noted by S_k $(k = 1, 2, \cdots, K)$ are generated by Equ.3. Through 1-D wavelet decompositions, the wavelet coefficients matrix of S_k is obtained, denoted by B_k. Using Equ.1 and Equ.2, all B_k are transformed into a 'large' binary matrix $G = \{g_1, g_2, \cdots, g_V\}$. By extending g_x to a row-vector f_x, all f_x are concatenated to constitute an ordered template feature vector to represent this class, denoted by $Ft = \{f_1, f_2, \cdots, f_V\}$.

For an input 2-D iris image, its feature extraction can be considered as a particular situation with $K = 1$ using the same threshold and the feature vector is denoted by $Fu = \{f'_1, f'_2, \cdots, f'_V\}$.

3.3 Iris Pattern Matching

The pattern matching between the feature vector of an unknown iris image and a template vector in an iris database is performed by a Hamming distance (HD) classifier, which can be calculated as:

$$HD = \frac{1}{\tilde{L}}\sum_{j=1}^{\tilde{L}} Fu_j \oplus Ft_j \tag{4}$$

where Fu and Ft are two binary vectors to be compared, $\tilde{L}$ is the length of the vector and $\oplus$ denotes the exclusive-OR operator. The result of this computation is then used as the quality of the match, with smaller values indicating better matches.

Due to changes of head orientation and binocular vergence at different iris acquisition time, the angular variation of an acquired iris image changes over a small range between -10 ° and +10 ° for the same person. Therefore, the following measures should be taken to make our proposed approach rotation invariant.

Firstly all the original iris images are normalized into rectangle blocks with size $H \times W'$ ($W' > W$). The centered blocks with size $H \times W$ from K representative iris samples are used to generate template feature vector. For an acquired iris image, the centered block with size $H \times W$ shifts from center to both sides by a step Δw, resulting in $(W' - W)/\Delta w$ overlapped blocks within the range of $W' - W$. We compute all the similarities between shifted blocks and the template vector. The minimum similarity is taken as the final matching score.

4 Experimental Results and Data Analysis

4.1 Determining Algorithm Parameters

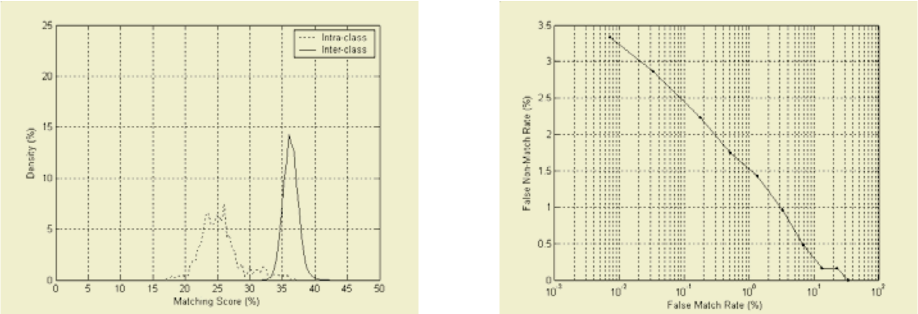
The Selection of the Parameters in 1-D Signal Generation. Each iris image is normalized into a rectangle block of a fixed size 48 × 512. Due to the appearance of some disturbing information such as eyelids or eyelashes in the lower part of the normalized image, the top-most 75% section is taken as the region of interest (ROI), in which we extract iris features. Since the number of rows in ROI is fixed, the product of the total number V of 1-D signals and the number U of rows used to form a 1-D signal is a constant in experiments. The recognition rate of the proposed algorithm can be regulated by changing the parameter U. A small U results in characterizing the iris details more completely, and thus increases recognition accuracy. A large U, however, implies a lower recognition rate with a higher computational efficiency. This way, we can trade off between speed and accuracy. In experiments, we chose $U = 4$ and in total 8 signals are generated from one iris image.

The Selection of the Wavelet Basis and Decomposition Scales. Due to desirable properties concerning compact support, vanishing moment, etc., the quadratic spline wavelet basis was adopted for 1-D multi-scale decompositions in this paper. For a signal with a certain length L, wavelet decompositions could decompose it over $\log_2 L$ scales with Mallat's pyramidal algorithm. With the scale increasing, the ability to represent a signal with wavelet coefficients turns degrading. Therefore, we used a combination of scales from 1 to 4 to construct wavelet coefficients matrices.

The Selection of the Binary Threshold T. The selection of the threshold T in Equ.2 directly affects the performance of our method. Because a combination of scales from 1 to 4 was selected and $\gamma \geq e^2$, we could compute that $T \in [0, 2.894]$ by the form $T = \sqrt{2(\ln \tilde{N} - \ln \gamma)}$. A larger T leads to more 0s in a feature vector, which weakens some useful information. A smaller T leads to more 1s, which magnifies some irrelevant information. Therefore, a reasonable T should be determined to perform the threshold-based scheme, which brings on appropriate feature representations. Therefore, we chose $T = 0.672$.

Table 1. Experiment results for different matching criteria.

Matching Criterion	FNMR (%)	FMR (%)
0.34	2.22	0.17
0.345	1.74	0.52
0.35	***1.42***	***1.34***
0.355	0.95	3.30
0.36	0.47	6.86



(a)

(b)

Fig. 4. Overall test results: (a) HD distributions of intra-class and inter-class; and (b) ROC curve.

4.2 Overall Test Results

Since the algorithm parameters were determined, we could use an iris database named CASIA [11] to evaluate the performance of the proposed method. In following experiments, we made in total 56,700 comparisons, including 630 intra-class comparisons and 56,070 inter-class comparisons. The receiver operating characteristic (ROC) curve and equal error rate (EER) are used to evaluate the performance [2]. Fig. 4 illustrated the HD distribution of intra-class and inter-class and the ROC curve. The performances of different matching criteria were tabulated in Table 1.

4.3 Data Analysis

(1) The distribution map in Fig. 4a illustrated that the HD distribution for intra-class irises was concentrated in the range $[0.22, 0.28]$. On the other hand, for inter-class irises, the HD distribution was concentrated in the range $[0.34, 0.39]$. This reveals that the features extracted by our method could meet the classification demand for iris recognition.

(2) Table 1 suggested that the EER was close to 1.40% (the italic and bold numbers in the table). From ROC curve in Fig. 4b, when the FMR was close to 0%, the FNMR was just less than 3.5%. In such experiments with less intra-class iris comparisons, the experimental results are highly encouraging.

5 Conclusions

This paper proposed a new iris recognition method based on quadratic spline wavelet multi-scale decompositions. Utilizing the probability theory and lemma, a threshold-based scheme is applied to the wavelet coefficients matrices to construct iris feature representations. The binary feature vector allows a simple pattern matching approach. The iris features can be compactly represented by 480 bytes and the average computation time for one iris recognition is less than one second. Therefore, the proposed iris recognition system fully meets the demands of information storage and real-time operation.

Acknowledgments

Portions of the research in this paper use the CASIA iris image database collected by the Institute of Automation, Chinese Academy of Sciences. We would like to thank them for generously providing the iris database.

References

1. Jain, A., Bolle, R. and Pankanti, S. Eds.: Biometrics: Personal Identification in a Networked Society. Norwell, MA: Kluwer, 1999.
2. Mansfield, T., Kelly, G., Chandler, D. and Kane, J.: Biometric Product Testing Final Report. Nat. Physical Lab., Middlesex, U.K., 2001.
3. Daugman, J.: High Confidence Visual Recognition of Persons by a Test of Statistical Independence. IEEE Trans. on PAMI. **vol.15, no.11** (1993) 1148–1161
4. Wildes, R.P. and Asmuth, J.C.: A Machine Vision System for Iris Recognition. Machine Vision Applicat. **vol.9** (1996) 1–8
5. Boles, W.W. and Boashash, B.: A Human Identification Technique Using Images of the Iris and Wavelet Transform. IEEE Trans. Signal Processing. **vol.46** (1998) 1185–1188
6. Zhu, Y., Tan, T. and Wang, Y.: Biometric Personal Identification Based on Iris Patterns. Proc. of IAPR, Int. Conf. Pattern Recognition (ICPR'2000). **vol.II** (2000) 805–808
7. Lim, S., Lee, K., Byeon, O. and Kim, T.: Efficient Iris Recognition through Improvement of Feature Vector and Classifier. ETRI Journal. **vol.23, no.2** (2001) 61–70
8. Ma, L., Wang, Y. and Zhang, D.: Efficient Iris Recognition by Characterizing Key Local Variations. IEEE Trans. Image Processing, **vol.13, no.6** (2004) 739–750
9. Pittner, S. and Kamarthi, S.V.: Feature Extraction from Wavelet Coefficients for Pattern Recognition Tasks. IEEE Trans. PAMI. **vol.21, no.1** (1999) 83–88
10. Ma, L.: Personal Identification Based on Iris Recognition. Ph.D dissertation, Inst. Automation, Chinese Academy of Sciences, Beijing, China, June 2003.
11. CASIA Iris Image Database. http://www.sinobiometrics.com.

Part VIII

Speech Recognition

An Utterance Verification Algorithm in Keyword Spotting System*

Haisheng Dai[1], Xiaoyan Zhu[2], Yupin Luo[1], and Shiyuan Yang[1]

[1] Department of Automation, Tsinghua University
100084 Beijing, China
dai@mail.au.tsinghua.edu.cn
{Luo,ysy-dau}@tsinghua.edu.cn
[2] Department of Computer Science and Technology, Tsinghua University
100084 Beijing, China
Zxy-dcs@tsinghua.edu.cn

Abstract. Speech Recognition and Verification are important components of a Keyword Spotting System whose performance is determined by both of them. In this paper, a new model-distance based utterance verification algorithm is proposed to improve the performance of the keyword spotting system. Furthermore, a substitution-error recovery module is built for improving of the system, which enhances the detection rate with the false-alarm rate remaining almost the same. Experiment shows that with the novel utterance verification method and new added substitution-error recovery module, the system performance is improved greatly.

1 Introduction

Keyword spotting system (KWS) is regarded as a branch of continuous speech recognition (CSR) because the goal of KWS is detecting the given keywords from unrestricted continuous speech signals. However, KWS is different from CSR because of the following reasons: (1) CSR has a limited vocabulary which restricts the searching space of CSR, while the vocabulary of KWS is not restricted; (2) utterance verification (UV) is essential in KWS, but is not necessary as a post-processing module in CSR. (3) In CSR, all the missed or recognized words are of same importance, while in KWS only the keywords are considered. Such that the evaluation measures for CSR are substitution-error rate, insertion-error rate and deletion-error rate, while those for KWS are detection rate and false alarm rate (FAR).

Keyword spotting system has been developed since 1970s. Christinansen et al [1] gave the definition of “keyword” and detected the keyword from continuous speech signals by using linear predictive coding (LPC). Later, Wohlford [2] proposed the concept of “filler” model. Wilpon et al [3] built a practical KWS based on hidden Markov models (HMM). Rohlicek et al [4] applied continuous HMM to KWS and proposed the evaluation criterion of keyword spotting system. Recently, two-pass based KWS methods become more and more mature [5]. In the first pass CSR is used

* This paper is supported by Natural Science Foundation No.60272019 and No. 60321002.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 555–561, 2005.

to generalize the hypothesis and then utterance verification is applied to verify it in the second pass. Lleida et al [6] integrated speech recognition and utterance verification into a one-pass procedure. Because CSR and utterance verification module play key roles in KWS, recent studies of KWS focus on how to improve the accuracy rate and robustness [7] of CSR and the performance of utterance verification algorithm [8-9].

Even CSR and utterance verification algorithm is improved to a certain level, there still exist other problems in keyword spotting system: (1) if a deletion-error occurs, the corresponding keyword is miss-detected, thus degrades the detection rate. (2) If a substitution-error occurs, besides miss-detected keywords, it brings false-alarm once the utterance verification procedure accepts the false hypothesis.

To solve the problem of deletion-error, SchWartz [10] constructs a N-Best word-lattice to expand the searching space which can increase the detection-rate; and Wu et al [11] recover the deleted keyword by building a partial pattern tree (PPT) of language model. However, up to now there is no effective technique to reduce the problem introduced by substitution-errors. In this paper, we aim to solve such problem by proposing a substitution-error recovery algorithm.

To decrease the false alarm rate, we described a model-distance based utterance verification algorithm, whose detailed description was proposed by Dai [9]. Furthermore, to improve the detection rate, a model-distance based substitution-error recovery algorithm is proposed to recover the substituted keywords from the rejected hypothesis in utterance verification procedure. However, the false alarm rate degrades at the same time. To solve this contradiction, a restricted substitution-error recovery algorithm is introduced. This algorithm is able to improve the detection rate remarkably while the false alarm rate remains at a low level.

The organization of the paper is as follows. Section 2 simply introduces a model-distance based utterance verification algorithm. Section 3 describes two kinds of substitution-error recovery algorithms: one is a non-restricted model-distance based algorithm, the other is a restricted one. Detailed experimental results are shown in section 4. Finally, we make our conclusion in section 5.

2 Model-Distance Based Utterance Verification Algorithm

In this section the definition and property of model-distance are introduced, and then model-distance based utterance verification algorithm is described.

2.1 Definition and Property of Model-Distance

Suppose that there are *N* Keywords ($M_1, M_2, \ldots, M_N$), and O_i is a perfect training data for M_i. For each M_j, the viterbi decoding procedure computes out the likelihood value ***P(O**$_i$**|M**$_j$**)***. The distance from M_j to M_i is defined as follows:

$$\mathrm{d}(\mathrm{M_j},\mathrm{M_i}) = \sum_{k=1}^{N} \mathrm{sgn}(p(O_i \mid M_k), p(O_i \mid M_j)) \tag{1}$$

where, *sgn(•)* is defined as:

$$\operatorname{sgn}(x,y)=\begin{cases}1 & ,x>y\\ 0 & ,x\le y\end{cases} \tag{2}$$

From the definition we can find that *d(Mj , Mi)* equals to the rank of M_j in the N-Best result while the decoded signal is the perfect training data of M_i. Apparently, less $d(M_{j\,,}\ M_i)$ is, more similar M_j and M_i are. And model-distance has following properties:

$$\mathrm{d}(\mathrm{M}_\mathrm{j},\mathrm{M}_\mathrm{i})\in\{0,1,......,N-1\} \tag{3}$$

$$\mathrm{d}(\mathrm{M}_\mathrm{i},\mathrm{M}_\mathrm{i})=0 \tag{4}$$

$$\sum_{i=1}^{N}d(M_j,M_i)=\frac{N(N-1)}{2} \tag{5}$$

We define model-distance matrix as

$$A_{NN}=[a_{ij}]_{NN} \tag{6}$$

$$a_{ij}=E[d(M_j,M_i)] \tag{7}$$

where, *E[•]* is mathematic expectation. While the vocabulary set is determined, the model-distance matrix can be trained by the training data.

In the following sections, this model-distance matrix is used for utterance verification and substitution-error recovery algorithms.

2.2 Model-Distance Based Verification Criterion

In the keyword spotting system, for an input signal O_i, we can get a hypothesis $\hat{M}_i$, from which $d(M_j,\hat{M}_i)$ is computed for each M_j, *j=1, 2, ..., N*. In our algorithm, the difference between the expectation of model-distance of real keyword a_{ij} with the model-distance of hypothesis $d(M_j,\hat{M}_i)$ is used to determine whether $\hat{M}_i$ is true or not, which can be formulated as:

$$\delta_i=\xi\sum_{j=1}^{N}\left|a_{ij}-d(M_j,\hat{M}_i)\right|, \tag{8}$$

where $\xi = \frac{2}{N^2}$ is a normalized coefficient which satisfies $0 \le \delta_i \le 1$ (this property is proved in the appendix of [9]). If δ_i is less than a given threshold, the hypothesis is accepted, otherwise rejected.

3 Substitution-Error Recovery Algorithm

In this section, model-distance based substitution-error recovery algorithm is introduced to improve the detection rate. Unfortunately, it increases the false-alarm rate at the same time. To solve this problem, restrictions such as relative likelihood value and syllable accuracy rate are imposed. Fig.1 shows the structure of keyword spotting system integrating the substitution-error recovery.

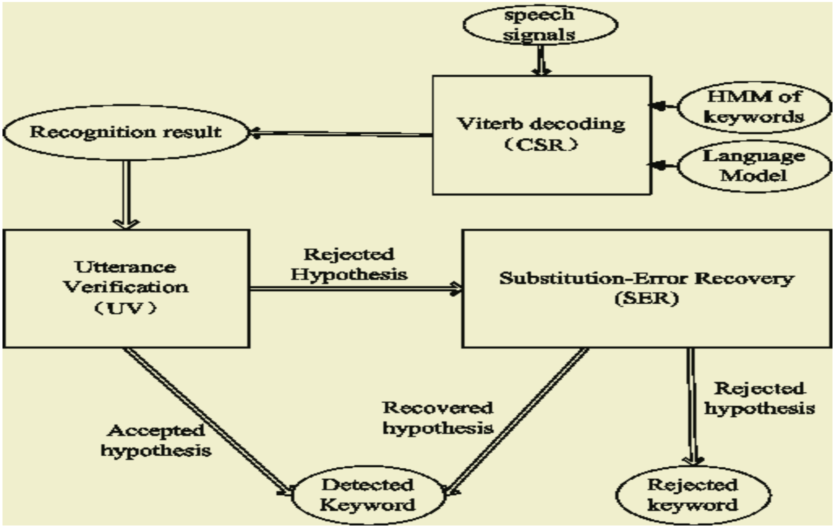


Fig. 1. Structure of KWS integrated substitution-error recovery procedure

3.1 Model-Distance Based Substitution-Error Recovery Algorithm

From Fig. 1 we can find that the hypothesis $\hat{M}_i$ is accepted by KWS if it is accepted by utterance verification procedure. In the other hand if it is rejected by utterance verification procedure, the corresponding signal O and the hypothesis $\hat{M}_i$ are processed by the following substitution-error recovery algorithm:

(1) For each model M_j, $p_j = P(O|M_j)$ and $d(M_j, \hat{M}_i)$ can be calculated, $d(M_j, \hat{M}_i)$ represents the rank of M_j in the N-Best recognition results;

(2) Calculate $\delta(M_k)$ by using model-distance matrix A_{NN}:

$$\delta(M_k) = \frac{2}{N^2}\sum_{j=1}^{N} | a_{kj} - d(M_j, \hat{M}_i) | \tag{9}$$

(3) Let k*= $\underset{k}{\operatorname{argmin}}(\delta(M_k))$. If $k^* \neq i$, M_{k^*} is considered as a substituted keyword and KWS accepts it as the detection result.

Eventually, the accepted hypothesis by utterance verification procedure and the recovered keyword by substitution-error recovery module are taken as the final detected keywords of keyword spotting system.

3.2 Restricted Substitution-Error Recovery Algorithm

The model-distance based substitution-error recovery algorithm recovers keywords from the rejected hypothesis, but inherently increases false-alarm rate. Thus, it is necessary to find out measures to control the recovering. For the hypothesis M_j, the following rules are applied:

(1) $d(M_j, \hat{M}_i) < T_1$, T$_1$ is a given threshold;

(2) $\frac{p(X | w_j)}{\max\{p(X | w_i)\}} > T_2$, T$_2$ is a given threshold;

(3) $\frac{N_{\mathrm{Rec}}}{N(M_j)} = 1$, N_{Rec} is the number of detected syllables, and $N(M_j)$ is the number of syllables in keyword M_j.

Only if all the above conditions are satisfied, M_j is recovered as a detected keyword.

4 Experiments

The platform is a continuous speech elevator control simulation system with a keyword spotting engine. The whole system consists of four components: speech recognition module, speech verification module, substitution-error recovery module and dialog management module.

The speech data is a corpus from spontaneous utterances of 10 people. There are totally 100 entities in the corpus, 50 of which is used as the keyword set, and others as words out of vocabulary. The speech feature vectors used in all of experiments are 39 dimensions, including 12 MEL-Frequency Cepstrum coefficient (MFCC), plus its 1st and 2nd order differences, as well as 1 normalized energy plus its 1st and 2nd order differences. The corresponding thresholds are determined by the experiments.

In Figure 2, Receiver Operating Characteristics (ROC) of several different utterance verification algorithms is presented. The details of the combination-score based utterance verification algorithm could be found in [8]. The experiment results show that our model-distance based utterance verification algorithm (real line) outperforms the combination-score based utterance verification algorithms. But the detection rate is still not high enough for a practical system.

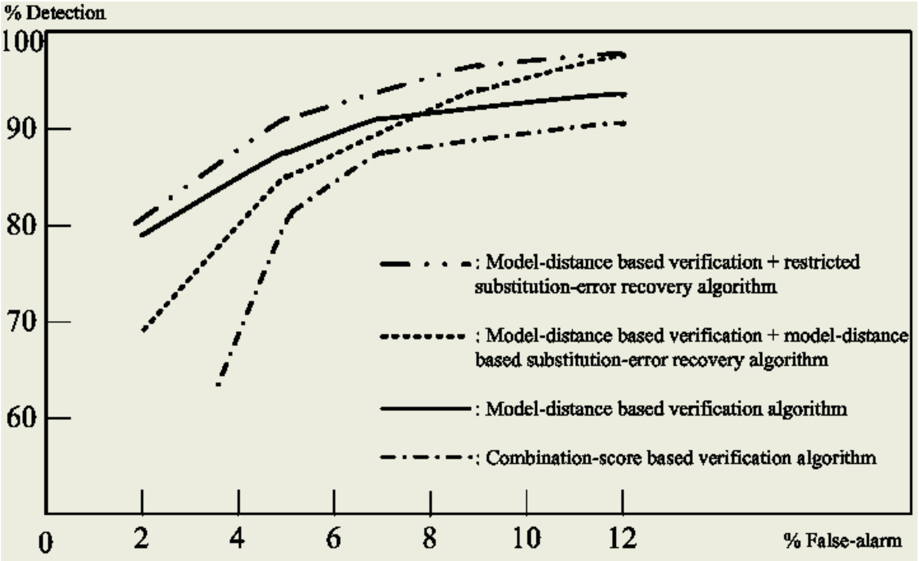


Fig. 2. The ROC curve of KWS for different utterance verification algorithms and integrated substitution-error recovery algorithms

By recruiting a model-distance based substitution-error recovery module, it is found that the curve of the detection performance changed in the whole range from lower false alarm to higher areas. We take the model-distance based utterance verification algorithm as a baseline. At the beginning the detection rate of the substitution-error recovery model is lower than that of the baseline. Finally, it can achieve a higher detection rate, however, with a higher false alarm rate. The reason for this is that too many rejected hypotheses were recovered incorrectly, resulting in a high false alarm rate. In contrast, by applying constraints on the recovering algorithm, the restricted substitution-error recovery algorithm overcomes the prior weakness. About 2% to 6% improvement of detection rate is obtained, comparing with our baseline. It was also found that, the restricted algorithm improves the detection performance over the combination-score based model by as much as 12%.

5 Conclusions

In this paper, a model-distance based utterance verification algorithm is proposed, which highly improves the performance of keyword spotting system. Furthermore, a model-distance based substitution-error recovery algorithm is given to enhance the detection rate. The experiment result shows that it increased the detection rate greatly with some added false alarms. To overcome this problem, a restricted substitution-error recovery algorithm is built to restrain some false recovered utterance signals. The new algorithm works well with a higher detection rate and almost same false alarm rate with baseline system.

References

1. Christiansen, R.W., Rushforth, C.K., “Detecting and Locating Key Words in Continuous Speech Using Linear Predictive Coding”, TEEE Trans. On ASSP, vol. ASSP-25, No.5, pp.361-367, Oct.1977
2. Higgins, Alan L., Wohlford, Robert E., “Keyword Recognition Using Template Concatenation”, ICASSP-85, vol.3 pp.1233-1236
3. Wilpon, J.G., Lee, C.H., Rabiner, L.R., “Application of Hidden Markov Models for Recognition of a Limited Set of Words in Unconstrained Speech”, ICASSP-89, vol.3, pp.254-257
4. Rohlicek, J. Robin, Russel, William, Roukos, Salim, Gish, Herbert, “Continuous Hidden Markov Modeling for Speaker-Independent Word Spotting”, ICASSP-89, vol.3, pp.627-630
5. Tadj,C., Poirier,F., “Keyword Spotting Using Supervised/Unsupervised Competitive Learning”, ICASSP-95, vol.1, pp.301-304
6. E. Lleida and R. C. Rose, “Utterance Verification in Continuous Speech Recognition: Decoding and Training Procedures,” IEEE Trans. on Speech and Audio Processing [J], 2000, Vol.8(2) pp. 126-139
7. Sirko Molau, Daniel Keysers, Hermann Ney, “Matching training and test data distributions for robust speech recognition”, Speech Communication 2003
8. Binfeng Yan, Rui Guo, Xiaoyan Zhu, Continuous Speech Recognition and Verification based on a Combination Score, EUROSPEECH 2003, Geneva
9. Dai Haisheng, Zhu Xiaoyan, Luo Yupin, Yang Shiyuan, “A Novel Keyword Verification Algorithm”, accepted by ACTA Electronica Sinica
10. R. SchWartz, “Efficient, High-Performance Algorithms for N-Best Search”, 1990 DARPA Speech and Natural Language Workshop, pp.6-11
11. Chung-Hsien Wu, Yeou-Jiunn Chen, “Recovery from false rejection using statistical partial pattern trees for sentence verification”, Speech Communication, 2003

A Clustering Algorithm for the Fast Match of Acoustic Conditions in Continuous Speech Recognition

Luis Javier Rodríguez and M. Inés Torres*

Pattern Recognition & Speech Technology Group
DEE. Facultad de Ciencia y Tecnología, Universidad del País Vasco
Apartado 644, 48080 Bilbao, Spain
luisja@we.lc.ehu.es

Abstract. In practical speech recognition applications, channel/environment conditions may not match those of the corpus used to estimate the acoustic models. A straightforward methodology is proposed in this paper by which the speech recognizer can match the acoustic conditions of input utterances, thus allowing instantaneous adaptation schemes. First a number of clusters is determined in the training material in a fully unsupervised way, using a dissimilarity measure based on shallow acoustic models. Then accurate acoustic models are estimated for each cluster, and finally a fast match strategy, based on the shallow models, is used to choose the most likely acoustic condition for each input utterance. The performance of the clustering algorithm was tested on two speech databases in Spanish: SENGLAR (read speech) and CORLEC-EHU-1 (spontaneous human-human dialogues). In both cases, speech utterances were consistently grouped by gender, by recording conditions or by background/channel noise. Furthermore, the fast match methodology led to noticeable improvements in preliminary phonetic recognition experiments, at 20-50% of the computational cost of the ML match.

1 Introduction

One of the most challenging issues posed by current applications of continuous speech recognition is the increased acoustic variability due to spontaneous speech, speaker features, channel or environmental conditions, etc. Many adaptation techniques have been proposed to increase the robustness of speech recognizers to speaker features and mismatched environment conditions [1]. One of them consists of organizing the training material into clusters of acoustically similar utterances, then training specific acoustic models for them, and finally matching the acoustic conditions (i.e. the most suitable cluster) for each input utterance.

* This work was partially supported by the University of the Basque Country, under grant 9/UPV 00224.310-13566/2001, and the Spanish MCYT, under project TIC2002-04103-C03-02.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 562–570, 2005.

The training material may be clustered in a supervised way by using *a priori* knowledge about speaker identities or environmental conditions of utterances [2]. But in practical applications such knowledge might be unreliable or unavailable. In this framework, an unsupervised clustering algorithm is needed to automatically determine an optimal partition in the set of utterances, as some authors have proposed [3–5].

In a previous study, we developed a clustering algorithm to find an optimal partition of speakers in the training material. Then we trained speaker-class models and during recognition the most suitable speaker classes for each input utterance were selected or combined in a fast and straightforward manner, using shallow acoustic models [6]. In that study, all the samples from any given speaker had to be moved to the same speaker-class. Here we apply the same methodology but in a fully unsupervised way: information about speaker identity is left out and each utterance is moved independently.

Assuming a non homogeneous set of speech utterances in the training corpus, we propose an unsupervised clustering algorithm which automatically finds an *optimal* partition in that set, using a dissimilarity measure based on shallow acoustic models. Once the optimal partition is defined, hidden Markov models (HMM) are estimated for each cluster. During recognition, the shallow models are applied to the input utterances in a straightforward manner, without recognizing them, to choose the most suitable clusters. The corresponding HMMs are then applied to get a number of decodings for each input utterance, and finally the most likely string is hypothesized. The number of decodings actually done depends on the *sharpness* of the decision, i.e. on the number of cluster candidates. In the best case, a single decoding would be carried out for each input utterance. Assuming that each cluster represents specific acoustic conditions (a pool of gender, channel and environment), this procedure can be viewed as a fast match of acoustic conditions. The fast match strategy is critical to making cluster models useful in actual applications, since the *Maximum Likelihood* (ML) match – i.e. carrying out all the decodings, one for each cluster, and then selecting the decoded string with the highest likelihood – would be too costly.

The rest of the paper is organized as follows: Section 2 describes the histogram models used to represent the clusters; Section 3 briefly outlines the clustering algorithm; Section 4 describes the fast match approach followed in this study; Section 5 presents experimental evaluation of the clustering algorithm on two speech databases in Spanish, and phonetic recognition results which provide evidence of the usefulness of the fast match strategy; finally, Section 6 summarizes the main contributions of the study.

2 The Shallow Acoustic Model

Let M be the number of acoustic vectors used to represent the speech signal at each time t. Then each sample $X(t)$ consists of M vectors, $X_j(t)$ with $j = 1, \ldots, M$. First, Vector Quantization (VQ) is applied to build a *codebook* of N centroids for each acoustic representation. These codebooks minimize the average distortion in quantifying the acoustic vectors of the training corpus. Once the VQ codebooks are defined, each vector $X_j(t)$ can be replaced by a single symbol $Y_j(t) \in \{1, \ldots, N\}$, corresponding to the index of the nearest centroid.

Now, assuming that the training corpus is partitioned into S clusters, consider the cluster i, for which $c(i)$ samples are available. We store in $c(k, j, i)$ the number of times $Y_j(t) = k$ in the set of samples corresponding to the cluster i, and define the discrete distribution $P_j(k|i)$ as:

$$P_j(k|i) = \frac{c(k,j,i)}{c(i)} \quad . \tag{1}$$

This is an empirical distribution based on the histograms of the symbols at each acoustic stream. Hereafter, we will refer to it as *histogram model.* Note that for any j $\sum_{k=1}^{N} c(k, j, i) = c(i)$, so that $\sum_{k=1}^{N} P_j(k|i) = 1$. The probability that a quantified speech sample $Y(t)$ is produced in the acoustic conditions represented by cluster i is defined as the joint discrete distribution:

$$P(Y(t)|i) = \prod_{j=1}^{M} P_j(Y_j(t)|i) \quad . \tag{2}$$

Finally, the probability that a speech utterance $Y = \{Y(t)|t = 1, \ldots, T\}$ is produced in the acoustic conditions represented by cluster i is computed as follows:

$$P(Y|i) = \prod_{t=1}^{T} P(Y(t)|i) \quad . \tag{3}$$

3 The Clustering Algorithm

A top-down clustering scheme was applied starting from a single cluster, iteratively splitting one of the clusters and readjusting the allocation of utterances until not enough speech frames were available or the average distortion decreased below a certain threshold.

Before writing the algorithm, we must give some definitions. First, a histogram model is constructed for each speech utterance l, based on the set of quantified samples corresponding to that utterance, with $\Upsilon(l) = \{Y(t)|t = 1, \ldots, s(l)\}$, $s(l)$ being the length of the utterance. Then the *dissimilarity* of l with regard to a given cluster i, $d(l; i)$, is defined as follows:

$$d(l;i) = -\log \left\{ \frac{P(\Upsilon(l)|i)}{P(\Upsilon(l)|l)} \right\} \quad , \tag{4}$$

where $P(\Upsilon(l)|\cdot)$ is computed as the joint probability of all the quantified speech samples corresponding to the utterance l, given a histogram model (equation 3).

At any iteration n of the clustering algorithm, each utterance l is assigned to the *closest* cluster $i_n^{(l)}$ in the partition $\Pi(n)$: $i_n^{(l)} = \arg\min_{g \in \Pi(n)} d(l; g)$. Taking this into account, the distortion of $\Pi(n)$ is defined as:

$$R(n) = \frac{1}{L}\sum_{l=1}^{L} d(l; i_n^{(l)}) = -log\left[\prod_{l=1}^{L}\frac{P(\Upsilon(l)|i_n^{(l)})}{P(\Upsilon(l)|l)}\right]^{1/L} \qquad (5)$$

where L is the number of speech utterances in the training corpus.

Finally, for each cluster i, the first and second centroid utterances, $\gamma_1^{(i)}$ and $\gamma_2^{(i)}$, are defined as those yielding the two smallest values of the dissimilarity with regard to that cluster:

$$\gamma_1^{(i)} = \arg\min_{l\in i} d(l;i) \;\; ; \quad \gamma_2^{(i)} = \arg\min_{l\in i, l\neq\gamma_1^{(i)}} d(l;i) \qquad (6)$$

The clustering algorithm is described in detail in the following paragraphs:

```
1. For each utterance $l \in \{1,\dots,L\}$ and for each acoustic stream $j \in \{1,\dots,M\}$, the *utterance histograms* $s(k,j,l)$ are counted, and the normalizing factor $s(l) = \sum_{k=1}^{N} s(k,1,l)$ computed.
2. Initially ($n = 0$), a single cluster is defined ($S = 1$) including all the utterances: $\forall l,\ i_0^{(l)} = 1$. The clustering distortion $R(0)$ is computed. Also, for each acoustic representation $j \in \{1,\dots,M\}$ the histogram model of the initial cluster is computed as follows: $c(k,j,1) = \sum_{l=1}^{L} s(k,j,l)$ and $c(1) = \sum_{l=1}^{L} s(l)$.
3. **repeat**
   3.1 $n \leftarrow n+1$
   3.2 For each cluster $g \in \Pi(n)$, obtain the first and second centroid utterances, $\gamma_1^{(g)}$ and $\gamma_2^{(g)}$, and the average cluster distortion, computed as $D(g) = \frac{1}{L(g)}\sum_{l\in g} d(l;g)$, where $L(g)$ is the number of speech utterances in $g$. Add this information to a list of *cluster split candidates*, $c_{cand}$, in descending order of $D(g)$.
   3.3 **while** $c_{cand} \neq \emptyset$ **do**
       3.3.1 Extract the first item of the list: $(g, \gamma_1^{(g)}, \gamma_2^{(g)})$, and split cluster $g$ in two, taking as seed models of the new clusters those of $\gamma_1^{(g)}$ and $\gamma_2^{(g)}$, respectively.
       3.3.2 **repeat**
                - For each utterance $l$, assign it to the nearest cluster
                - For each cluster $i$, recompute the histogram model using the counts $s(k,j,l)$ and $s(l)$ of the utterances assigned to it.
             **until** maximum number of iterations **or** clusters unchanged
       3.3.3 **if** the new partition is valid **then**
             { $S \leftarrow S+1$; Compute $R(n)$; Empty $c_{cand}$; }
             **else**
             { Recover the partition at $n-1$; $R(n) \leftarrow R(n-1)$; }
   **until** $(R(n-1) - R(n))/R(n) < \tau$
4. Store the partition information and the corresponding histogram models.
```

In the above algorithm $\tau > 0$ is an empirical threshold for the relative decrease in average distortion. Also, each time a new partition is generated, all the clusters must contain a minimum number of speech frames to guarantee the trainability of the acoustic models. When not enough frames are available for any of the clusters, the previous partition is recovered and another splitting explored (step 3.3.3). Note also that the candidate splittings are explored in descending order of $D(g)$, so that the cluster with the highest distortion is split first.

4 The Fast Match Strategy

During recognition, the most suitable acoustic model(s) must be selected/combined for each input utterance. Various alternatives were explored in a previous study, where each cluster represented a speaker class [6]. The *Maximum likelihood* (ML) match approach, consisting of carrying out S decodings, one for each HMM set, and selecting the one that yielded the highest likelihood, was found to be the optimal but also the most expensive alternative. On the other hand, if the histogram models were used to *pre-select a beam of candidates* – thus drastically reducing the number of decodings –, the same performance was obtained at a much lower cost. In practice, the average number of decodings was reduced to around two or three.

Taking these results into account, for each input utterance we have considered only those clusters whose histogram probabilities are higher than a heuristically fixed threshold (70% of the maximum value). Decodings are obtained only for them, and finally the decoded string that yields the highest likelihood is hypothesized. This is a kind of beam selection, motivated by the fact that sometimes the most suitable cluster – in terms of acoustic likelihood – yields histogram probabilities near but below the maximum.

5 Experimental Results

5.1 Databases

A phonetically and gender-balanced read speech database in Spanish, called SENGLAR, acquired at 16 kHz in laboratory conditions, was considered in the first place to tune the clustering algorithm. The training corpus consisted of 1529 utterances, pronounced by 57 (29 male, 28 female) speakers, and included 60399 phone samples with a total duration of around 80 minutes. The test corpus consisted of 700 utterances, pronounced by 33 (18 male, 15 female) speakers, and included 32034 phones with a total duration of around 40 minutes.

A spontaneous speech database in Spanish called CORLEC-EHU-1 [7], composed of 42 human-human dialogues taken from radio and TV broadcasts using an analog tape recorder, was considered in the second place to test the proposed methodology in more difficult conditions: variable and noticeable background/channel noise, presence of spontaneous speech events, pronunciation variability, etc. The training corpus consisted of 1421 utterances, pronounced by 67 (49 male, 18 female) speakers, and included 187675 phone samples with a total duration of around 225 minutes. The test corpus consisted of 704 utterances, pronounced by 35 (21 male, 14 female) speakers, and included 93415 phones with a total duration of around 114 minutes.

5.2 Results of Clustering

The mel-scale cepstral coefficients (MFCC) and energy (E) – computed in frames of 25 milliseconds, taken each 10 milliseconds – were used as acoustic features. The first and second derivatives of the MFCCs and the first derivatives of E were also computed. Four acoustic streams were defined: MFCC, ΔMFCC, Δ^2MFCC and (E,ΔE). Finally, the LBG vector quantization algorithm [8] was applied to get four codebooks, each one consisting of 256 centroids.

The clustering algorithm was run using the training corpora of the two databases described above. At least 30000 speech frames (5 minutes) were required for each cluster to be valid. The maximum number of convergence iterations (step 3.3.2) was set at 20, and the threshold for the relative decrease in average distortion was set at $\tau = 0.01$. This resulted in 8 clusters for SENGLAR and 17 clusters for CORLEC-EHU-1.

SENGLAR was built by integrating three sub-corpora, called *FRASES*, *EUROM1* and *PROBA*, recorded in different places with slightly different hardware, so that not only speaker characteristics but also channel features may differ from one utterance to other. As shown in Table 1, all the clusters except for #3 and #4 consisted of utterances from one single sub-corpus. Additionally, clusters were formed almost exclusively either by male or by female speakers. This means that channel and speaker characteristics were effectively working to separate clusters from one another.

With regard to CORLEC-EHU-1, besides gender, two channel/environment conditions were clearly separated by the clustering algorithm: radio and TV interviews. In fact, 13 of the 17 clusters were pure in terms of gender and channel/environment, which represents 51.76% of the training frames. The remaining 4 clusters consisted of a pool of male/female, radio/TV utterances.

Table 1. Distribution of speech utterances after clustering in SENGLAR.

	FRASES		**EUROM1**		**PROBA**	
	male	**female**	**male**	**female**	**male**	**female**
Cluster #1	0	0	120	8	0	0
Cluster #2	119	0	0	0	0	0
Cluster #3	0	0	302	2	60	0
Cluster #4	0	0	1	6	14	100
Cluster #5	0	0	24	236	0	0
Cluster #6	0	262	0	0	0	0
Cluster #7	0	0	0	143	0	0
Cluster #8	132	0	0	0	0	0

5.3 Phonetic Recognition Results

Phonetic recognition experiments were carried out using the HMMs obtained through the unsupervised clustering methodology described above. MAP estimates were applied to get more robust models (only the Gaussian means and weights were re-estimated) [9]. During recognition, the fast match strategy described in Section 4 was applied. In the case of SENGLAR, the set of context-independent sublexical units consisted of 23 phone-like units (PLUs) plus one extra unit for *silences*. In the case of CORLEC-EHU-1, besides the 23 PLUs 14 extra units were defined to model spontaneous speech events such as noises, lengthenings, filled pauses, silent pauses, etc. A set of left-side biphones was also defined in both cases, taking into account only the trainability of the corresponding models (at least 300 training samples were required). Left-side biphones were applied jointly with context-independent units to guarantee acoustic coverage. Each sublexical unit was represented with a left-right Continuous-Density HMM consisting of three states with self-loops but no skips. Phonological restrictions were applied only when dealing with left-side biphones. Finally, the extra units representing spontaneous speech events were either filtered or mapped into PLUs before the recognized and the correct strings were aligned. Phonetic recognition rates obtained using HMMs adapted through unsupervised clustering are shown in Table 2. To allow suitable comparisons, results using non-adapted HMMs (estimated using the whole training corpus) and HMMs adapted through speaker clustering [6] are also shown.

Table 2. Phonetic recognition rates obtained using non-adapted HMMs and HMMs adapted through speaker clustering and unsupervised clustering of utterances, for SENGLAR and CORLEC-EHU-1. Experiments were carried out using context-independent (CI) and context-dependent (CD) sublexical units.

	SENGLAR		**CORLEC-EHU-1**	
	CI	**CD**	**CI**	**CD**
Non-adapted HMMs	72.38	75.38	52.42	57.09
Adapted HMMs: Speaker Clustering	74.41	75.79	53.89	58.05
Adapted HMMs: Unsupervised Clustering	**74.33**	**75.78**	**53.53**	**57.58**

The HMMs adapted through unsupervised clustering outperformed the non-adapted HMMs in all cases. In the case of SENGLAR, improvements were quite noticeable when using context-independent models (7.06% relative error reduction), whereas only slight imoprovements were achieved with context-dependent models (1.62% relative error reduction). This is probably due to a lack of samples for the context-dependent models. In the case of CORLEC-EHU-1 more training samples were available, but the higher acoustic variability of spontaneous speech and especially the adverse channel/environment conditions made the improvements smaller in both cases (2.33% and 1.17% relative error reduction, respectively). In fact, phonetic recognition rates for CORLEC-EHU-1 are around 20 absolute points lower than those obtained for SENGLAR. So, though the usupervised clustering of utterances helps in modeling channel/environment variabilities, more specific strategies (noise compensation techniques, noise robust features, etc.) seem to be needed. On the other hand, the performance attained through unsupervised clustering is almost the same as that obtained through speaker clustering, with no information about either speaker identities or channel/environment conditions. Finally, the average number of decodings in the fast match was 4.09 in the case of SENGLAR and 3.64 in the case of CORLEC-EHU-1, which works out at 51.13% and 21.41% of the computational cost of the ML match, respectively.

6 Concluding Remarks

A new clustering algorithm is presented in this paper which automatically determines an optimal partition in the training corpus of a speech database using a dissimilarity measure based on shallow acoustic models. Then accurate acoustic models are estimated for each cluster, which represent specific (but unknown) speaker/environment conditions. During recognition, the most suitable clusters are selected using a fast match strategy, combining acoustic probabilities computed with the shallow models and full decodings obtained with HMMs. Preliminary results are presented for two databases of read and spontaneous speech in Spanish, revealing that speaker and channel/environment characteristics are implicitly taken into account by the clustering algorithm. A 7% decrease in error rate was attained in phonetic recognition experiments over read speech, at half the computational cost of the ML match. For spontaneous speech, the relative error decrease was slightly higher than 2%, at 20% of the cost of the ML match. Our current work involves applying this methodology to larger corpora of non homogeneous speech, such as those recorded in human-machine dialogue tasks. Note that unsupervised adaptation to speaker and environment conditions is crucial to increasing the robustness of spoken dialogue systems.

References

1. Gales, M.J.F.: Adaptive Training for Robust ASR. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Madonna di Campiglio (Italy) (2001)
2. Gao, Y., Padmanabhan, M., Picheny, M.: Speaker Adaptation Based on Pre-Clustering Training Speakers. In: Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH). (1997) 2091–2094
3. Jin, H., Kubala, F., Schwartz, R.: Automatic Speaker Clustering. In: Proceedings of the DARPA Speech Recognition Workshop, Chantilly, VA (1997) 108–111
4. Chen, S.S., Gopalakrishnan, P.S.: Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA (1998)

5. Ajmera, J., Wooters, C.: A Robust Speaker Clustering Algorithm. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), St. Thomas, U.S. Virgin Islands (2003)
6. Rodríguez, L.J., Torres, M.I.: A Speaker Clustering Algorithm for Fast Speaker Adaptation in Continuous Speech Recognition. In Sojka, P., Kopeček, I., Pala, K., eds.: Proceedings of the 7th International Conference on Text, Speech and Dialogue (TSD 2004). Lecture Notes in Artificial Intelligence LNCS/LNAI 3206, Brno, Czech Republic, Springer-Verlag (2004) 433–440
7. Rodríguez, L.J., Torres, M.I.: Annotation and Analysis of Acoustic and Lexical Events in a Generic Corpus of Spontaneous Speech in Spanish. In: Proceedings of the ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo Institute of Technology, Tokyo, Japan (2003) 187–190
8. Linde, Y., Buzo, A., Gray, R.M.: An Algorithm for Vector Quantizer Design. IEEE Transactions on Communications **28** (1980) 84–95
9. Gauvain, J.L., Lee, C.H.: Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. IEEE Transactions on Speech and Audio Processing **2** (1994) 291–298

Adaptive Signal Models for Wide-Band Speech and Audio Compression

Pedro Vera-Candeas[1], Nicolás Ruiz-Reyes[1], Manuel Rosa-Zurera[2],
Juan C. Cuevas-Martinez[1], and Francisco López-Ferreras[2]

[1] Electronics and Telecommunication Engineering Department, University of Jaén
Polytechnic School, C/ Alfonso X el Sabio 28, 23700 Linares, Jaén, Spain
{pvera,nicolas,jccuevas}@ujaen.es

[2] Signal Theory and Communications Department, University of Alcalá
Polytechnic School, 28871 Alcalá de Henares, Madrid, Spain
{manuel.rosa,francisco.lopez}@uah.es

Abstract. This paper deals with the application of adaptive signal models for parametric speech and audio compression. The matching pursuit algorithm is used for extracting sinusoidal components and transients in audio signals. The resulting residue is perceptually modelled as a noise like signal. When a transient is detected, psychoacoustic-adapted matching pursuits are accomplished using a wavelet-based dictionary followed of an harmonic one. Otherwise, matching pursuit is applied only to the harmonic dictionary. This multi-part model (Sines + Transients + Noise) is successfully applied for speech and audio coding purposes, assuring high perceptual quality at low bit rates (close to 16 kbps for most of the signals considered for testing).

1 Introduction

Parametric coding of audio signals has become a popular tool for representing these signals at very low bit rates [1–3]. A wide range of audio signals intuitively fit into the three-part model of Sines, Transients and Noise. Transients describe drum hits and the stacks of many instruments, sines describe signal components that have a distinct pitch, and noise often describes the rest of the signal that is neither sinusoidal nor transient. This model consists of three parts that work together and complement each other to form a complete and robust signal model, which makes possible a highly optimized audio compression scheme. To alleviate model mismatch problems, the three part of the model operate in series. First, transients are modelled and removed, leaving a residual signal. Then, sinusoids are modelled and removed, leaving a noise-like signal for the noise model. As such, each model captures signal components that are coherent to its underlying assumptions.

The classical sinusoidal or harmonic model has been applied with success for the purpose of coding speech signals [4]. This model comprises an analysis-synthesis framework that represents a signal as the sum of a set of sinusoids (partials) with time-varying frequencies, phases, and amplitudes. A large number

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 571–578, 2005.

of methods have been proposed for estimating the parameters of the sinusoidal model. Estimation of parameters is typically accomplished by peak picking the Short-Time Fourier Transform (STFT) [4]. Usually, analysis by synthesis is used in order to verify the detection of every spectral peak.

On the other hand, transients extraction is useful for those parts of audio signals with sharp attacks, because sinusoidal and noise models cannot represent them efficiently. In [3, 5, 6] different approaches for transient modelling are presented.

The three-part signal model is completed with a noise model for noise-like signals. Noise modelling has seen attention in the literature. LPC based schemes are the subject of much research. Another promising noise model has perceptual roots in that it uses energy on an Equivalent Rectangular Bandwidth (ERB) scale [7]. In this paper the three-part signal model is completed with a wavelet-based noise model.

This paper proposes an efficient, accurate and flexible multi-part model for wide-band speech and audio coding. The matching pursuit algorithm is used in order to iteratively select the functions that best match the current audio frame for representing transients and sinusoids. Sinusoids are modelled using sets of complex exponential functions, while transients are modelled using sets of wavelet functions. The matching pursuit algorithm operates with both sinusoids and wavelet functions.

2 Matching Pursuit

The matching pursuit algorithm was introduced by Mallat and Zhang in [8]. So as to explain the basic ideas concerning this algorithm, let's suppose a linear expansion approximating the analyzed signal $x[n]$ in terms of functions $g_i[n]$ chosen from a over-complete dictionary $D = \{g_i \ ; \ i = 0, 1, \ldots, L\}$. The L elements of the dictionary span C^L and are restricted to have unit norm.

At the first iteration of matching pursuit, the atom $g_i[n]$ which gives the largest inner product with the analyzed signal $x[n]$ is chosen. The contribution of this vector is then subtracted from the signal and the process is repeated on the residue. At the m-th iteration, the residue is:

$$r^m[n] = \begin{cases} x[n] & m = 0 \\ r^{m+1}[n] + \alpha_{i(m)} \cdot g_{i(m)}[n] & m \neq 0 \end{cases} \tag{1}$$

where $\alpha_{i(m)}$ is the weight associated to the optimum atom $g_{i(m)}[n]$ at the m-th iteration, and $i(m)$ the dictionary index of that atom.

By computing the orthogonal projections of residue $r^m[n]$ on elements $g_i[n] \in D$, the weight associated to each element at the m-th iteration is got:

$$\alpha_i^m = \frac{\langle r^m[n], g_i[n] \rangle}{\langle g_i[n], g_i[n] \rangle} = \frac{\langle r^m[n], g_i[n] \rangle}{\|g_i[n]\|^2} = \langle r^m[n], g_i[n] \rangle \tag{2}$$

The l^2 norm of $r^{m+1}[n]$ can be expressed as:

$$||r^{m+1}||^2 = ||r^m||^2 - |\langle r^m, g_i \rangle|^2 = ||r^m||^2 - |\alpha_i^m|^2 \tag{3}$$

which is minimized by maximizing $|\alpha_i^m|^2 = |\langle r^m, g_i \rangle|^2$.

Therefore, the optimum atom $g_{i(m)}$ at the m-th iteration is obtained as:

$$g_{i(m)} = \arg \min_{g_i \in D} \|r^{m+1}\|^2 = \arg \max_{g_i \in D} |\alpha_i^m|^2 \tag{4}$$

It is simply equivalent to choosing the atom whose inner product with the signal has the highest value.

The computation of correlations $\langle r^m[n], g_i[n] \rangle$ for all $g_i[n] \in D$ at each iteration is highly computational consuming. As derived in [8], this computation effort can be substantially reduced using an updating formula based on equation (1). The correlations at the m-th iteration are given by:

$$\langle r^{m+1}[n], g_i[n] \rangle = \langle r^m[n], g_i[n] \rangle - \alpha_{i(m)} \cdot \langle g_{i(m)}[n], g_i[n] \rangle \tag{5}$$

where the only new computation required for the correlation updating procedure refers to the cross-correlation term $\langle g_{i(m)}[n], g_i[n] \rangle$, which can be pre-calculated and stored, once overcomplete set D has been determined.

3 The Proposed Wide-Band Speech and Audio Coder

The proposed parametric wide-band speech and audio coder is defined with three meaningful components:

- Transient modelling using energy-adaptive matching pursuit with a dictionary of wavelet functions.
- Sinusoidal modelling using psychoacoustic-adaptive matching pursuit with a dictionary a complex exponentials.
- Residue modelling as a noise like signal.

Figure 1 shows the encoder stage of the proposed parametric wide-band speech and audio coder.

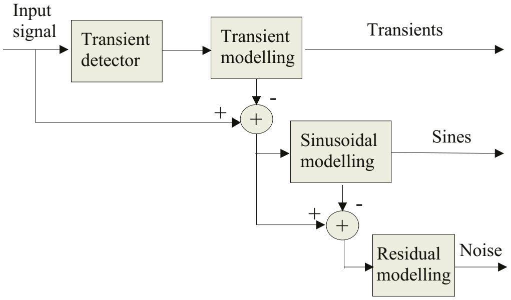


Fig. 1. Block diagram of the encoder stage.

The proposed wide-band speech and audio coder extracts from the input audio signal a set of different parameters to be sent to the decoder. These parameters represent the information provided by the three-part model (Sines + Transient + Noise). They are quantified using psycho-acoustical information to ensure that decoded signals are perceptually identical to the original ones.

Before transient modelling, transient detection is required. Our transient detector is based on sudden energy change detection. Besides, an adaptive tiling of the time axis is required to achieve a right performance of the proposed audio coder. We have used the algorithm proposed in [9].

3.1 Transient Modelling

We propose using matching pursuits with a dictionary of orthogonal wavelet functions for transient modelling. The overcomplete dictionary D is made up with those functions which give rise to the J-depth full Wavelet-Packet (WP) decomposition, being $M_{WP} = J \cdot N$ the WP dictionary size, and N the frame length. The inner products of the signal with the wavelet-based atoms in set D lead to all the wavelet coefficients that can be considered in the J-depth full WP tree. These coefficients can be identified using three indexes, $\{i, j, k\}$, which indicate the sub-band at a given decomposition depth, the decomposition depth and the delay, respectively. The wavelet coefficients at the m-th iteration of matching pursuit and the wavelet-based atoms can be expressed as follows:

$$\alpha^m_{\{i,j,k\}} = \langle r^m[n], g_{\{i,j,k\}}[n] \rangle \tag{6}$$

$$g_{\{i,j,k\}}[n] = g_{\{i,j\}}[n - 2^j k] \tag{7}$$

According to (5), the only necessary correlations to implement the matching pursuit are $\langle x[n], g_{\{i,j,k\}}[n] \rangle$ and $\langle g_{\{i_1,j_1,k_1\}}[n], g_{\{i_2,j_2,k_2\}}[n] \rangle$. The first ones are obtained from the WP transform of $x[n]$, while correlations between atoms are pre-calculated and memory stored. These cross-correlations are formulated in [6] when wavelet-based dictionaries built from orthonormal wavelets are used, which results in:

$$\langle g_{\{i_1,j_1,k_1\}}[n], g_{\{i_2,j_2,k_2\}}[n] \rangle = \begin{cases} \delta[k_2 - k_1] & i_1 = i_2, j_1 = j_2 \\ 0 & i_2 \neq \lfloor \frac{i_1}{2^{j_1 - j_2}} \rfloor \\ g_{\{i,j,k_1\}}[k_2] & i_2 = \lfloor \frac{i_1}{2^{j_1 - j_2}} \rfloor \end{cases} \tag{8}$$

where $j = j_1 - j_2$ and $i = ((i_1))_{2^j}$. Therefore, according to (8), the iterative procedure to update correlations requires impulsive responses of the synthesis WP tree branches to be stored [6].

3.2 Sinusoidal Modelling

For sinusoidal modelling, we propose using matching pursuits with a dictionary of windowed complex exponential functions, instead of a set of windowed sinusoidal functions, in order to reduce the computational complexity. Using windowed complex exponential sets, only the frequency of every exponential function must be determined, which involves a significant reduction of the dictionary size [10]. The functions that belong to the considered set can be expressed as follows:

$$g_i[n] = S \cdot w[n] \cdot e^{j\frac{2\pi i}{2L}n}, \quad i = 0, \ldots, L \tag{9}$$

The constant S is selected in order to obtain unit-norm functions, $w[n]$ is the N-length analysis window, and L+1 the number of frequencies within the dictionary. Amplitude, frequency and phase are the three parameters that define each extracted tone by the sinusoidal model.

The implemented matching pursuit algorithm for sinusoidal modelling is psychoacoustic-adaptive as in [11]. According to this approach, the extracted tone at each iteration is the perceptually most important one. Psychoacoustic-adaptive matching pursuits [11] define a perceptual distortion measure as

$$\|PD_i\|^2 = \int_0^1 \hat{a}(f)|(w[n](\widehat{\alpha_i^m} g_i[n]))(f)|^2\, df \tag{10}$$

where $\hat{}$ indicates the Fourier transform, $w[n]$ is a window defining the signal segment, and $\hat{a}$ the inverse of the masking threshold, which is computed on the basis of the reconstructed signal that changes at each iteration.

In our implementation, the perceptual distortion measure in equation (10) is slightly modified by integrating directly along the bark scale, which results in a complexity reduction.

3.3 Residual Modelling

After sinusoidal and transient modeling, the residue is considered to be a noise like signal. For audio applications, psychoacoustic phenomena have to be incorporated into the noise model. For noise perception, the exact shape of the magnitude spectrum is not as crucial as the energy at each critical band. According to this principle, the ERB noise modelling is proposed in [7]. In our approach, the ERB model is approximated by the Discrete Wavelet Transform (DWT). In this case, DWT dictates the form of the filter bank, performing a dyadic partition in frequency, which plays a central role in many aspects of perception.

The proposed noise model is composed of two stages: analysis and synthesis. The DWT-based analysis stage divides each frame into $J + 1$ wavelet bands (being J the decomposition depth), and estimates their energy. For the l-th frame, the energy of the r-th wavelet band is found as:

$$E_r^l = \sum_{m \in \beta_r} |X^l(m)|^2 \tag{11}$$

where β_r contains the indexes of the r-th wavelet band, and $X^l(m)$, $m \in \beta_r$, represents the wavelet coefficients of the r-th wavelet band for the l-th frame.

The energy parameters approximates a power spectrum with piecewise constant energy according to the DWT filter bank. These parameters are used for the DWT-based synthesis stage. In the synthesis stage the wavelet coefficients are initialized to white noise using each band energy to control its respective gain, which results in the synthesized noise. Subjective listening tests pointed the necessity of improving the time characteristics of the synthesized noise in order to avoid spreading effects. LPC filtering has been included in the proposed noise model to achieve a time shaping of the synthesized noise. We have applied an Auto-Regressive all poles model with 4 poles as maximum. The number of poles in the model is given by the prediction gain. A lattice structure is adopted to achieve an efficient quantization of the AR model information included in our noise modelling approach.

4 Results and Discussion

To assess the performance of the proposed wide-band speech and audio coder, we have obtained some subjective and objective results. The configuration parameters are: 32-coefficient Daubechies filters and 4-level full WP decompositions (J = 4) for transient modelling, 4096 frequencies (L = 4096) within the dictionary for sinusoidal modelling, and 32-coefficient Daubechies filters and 9-level depth for DWT in noise modelling. Twelve music samples considered hard to encode have been used. They are 15 seconds-length CD-quality one channel speech and audio signals. Special attention has been paid to signals with impulsive energy bursts, which are extremely susceptible to the presence of 'pre-echoes', and we have made sure that the chosen set of source material covers a wide variety of signals.

4.1 Objective Results

The resulting binary rates obtained with the proposed wide-band speech and audio coder are presented in table 1. It contains the partial bit rates resulting for the synthetic signals obtained from sinusoidal, transient and residual modelling and the final bit rates resulting for the decoded signals (in kbits/s).

In order to illustrate the performance of the proposed wide-band speech and audio coder, let's consider an audio frame with an impulsive energy burst. Figure 2(a) represents the original audio signal, while figures 2(b) and 2(c) represent the synthesized transient and sinusoidal components, respectively, when they are modelled using the above described approaches. Finally, figure 2(d) shows the noise-like residual signal. It can be observed that the synthetic signal in figure 2(b) properly represents the sharp attack in the original one.

4.2 Subjective Results

The subjective tests have been performed on headphones under the A-B-C rule using the twelve sequences shown in table 1. The A-B-C methodology, known as a triple-stimulus double blind test with hidden reference, is recommended by ITU-R in the BS. 1116-1 recommendation. Tests have been carried out with twenty trained listeners, and the results are shown in table 2.

Table 1. Bit rates.

Item	Description	Tones	Transients	Residue	Decoded signals
es01	Suzanne Vega	12.14	0.98	3.34	16.52
es02	German male speech	12.48	0.78	3.37	16.69
es03	English female speech	13.94	0.97	3.00	17.98
si01	Harpsichord	11.73	0.25	2.54	14.60
si02	Castanets	11.84	4.30	2.38	18.61
si03	Pitch pipe	8.21	0.15	3.50	11.90
sm01	Bagpipes	9.22	0.17	3.75	13.20
sm02	Glockenspiel	3.76	0.67	2.36	6.85
sm03	Plucked strings	13.94	0.14	2.80	16.93
sc01	Trumpet solo and orchestra	13.00	0.45	2.87	16.38
sc02	Orchestra piece	12.76	0.20	2.25	15.26
sc03	Contemporary pop	15.60	0.21	2.85	18.73

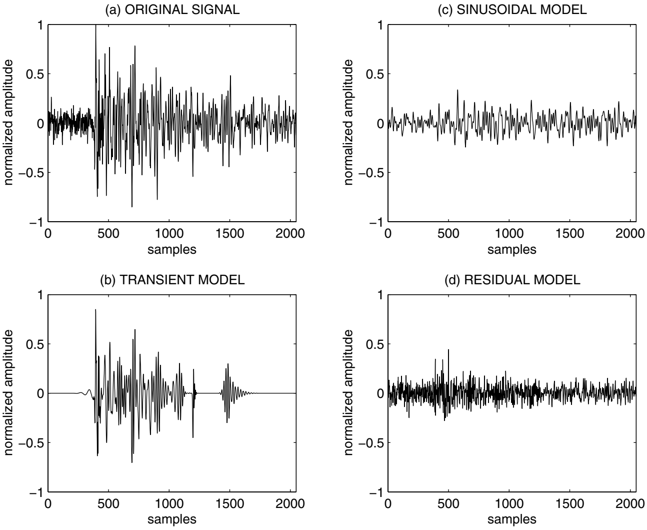


Fig. 2. Synthetic signals obtained from transient, sinusoidal and residual modelling.

5 Conclusions

This paper deals with parametric representation for wide-band speech and audio coding. The used model considers the speech and audio signals composed of three kinds of components: sinusoidal, transients and noise like components. For estimating the parameters of the sinusoidal and transient models, matching pursuit with dictionaries of complex exponentials and wavelet functions, respectively, is used. A novel wavelet-based noise modelling is applied for residue modelling, which is completed with LPC filtering to achieve Time Noise Shaping (TNS). The proposed wide-band speech and audio coder achieves nearly transparent coding

Table 2. Subjective results under the ITU-R BS.1116-1 recommendation.

Test Items	Orig. MOS	Decoded MOS	ΔMOS
es01	5.00	4.19	0.81
es02	5.00	4.02	0.98
es03	5.00	4.12	0.88
si01	4.97	4.63	0.34
si02	5.00	4.55	0.45
si03	5.00	4.33	0.67
sm01	4.99	4.51	0.48
sm02	4.98	4.68	0.30
sm03	4.97	4.75	0.22
sc01	5.00	4.40	0.60
sc02	5.00	4.28	0.72
sc03	5.00	4.33	0.67

at very low bit rates (close to 16 kbit/seg). Hence, our coder is a good proposal for audio coding applications at very low bit rates, as Internet streaming.

References

1. Levine, S., Smith, J.: A Sines+Transients+Noise Audio Representation for Data Compression and Time/Pitch Scale Modifications, *105th AES Convention*, preprint 4781 (1998).
2. Verma, T.S.: A perceptually based audio signal model with application to scalable audio compression, *PhD Thesis*, Standford University (1999).
3. Den Brinker, A.C., Schuiijers, A.G.P., Oomen, A.W.J.: Parametric coding for high quality audio, *112th AES Convention*, Preprint 5554 (2002).
4. McAulay, R., Quatieri, T.: Speech Analysis/Synthesis Based on a Sinusoidal Representation, *IEEE Trans. Acoustic, Speech and Signal Processing* **34** 4 (1986) 744-754.
5. Nieuwenhuijse, J., Heusdens, R., Deprettere, E.F.: Robust exponential modeling of audio signals, *Proc. ICASSP-98* **6** (1998) 3581-3584.
6. Vera-Candeas, P., Ruiz-Reyes, N., Rosa-Zurera, M., Martinez-Muñoz, D., Lopez-Ferreras, F.: Transient Modeling by Matching Pursuits with a Wavelet Dictionary for Parametric Audio Coding, *IEEE Signal Processing Letters* **11** 3 (2004) 349-352.
7. Goodwin, M.: Residual modelling in music analysis-synthesis, *Proc. ICASSP-96* **2** (1996) 1005-1008.
8. Mallat, S., Zhang, Z.: Matching pursuits with time-frequency dictionaries, *IEEE Trans. on Signal Processing* **41** (1993) 3397-3415.
9. Ruiz, N., Rosa, M., López, F., Vera, P.: New algorithm for achieving an adaptive tiling of the time axis for audio coding purposes, *Electronic Letters* **80** (2002) 434-435.
10. Goodwin, M.M.: Adaptive Signal Models. Theory, Algorithms and Audio Applications, *Kluwer Academic Publishers* (1998).
11. Heusdens, R., Vafin, R., Kleijn, W.B.: Sinusoidal Modelling using Psychoacoustic-Adaptive Matching Pursuits, *IEEE Signal Processing Letters* **9** 8 (2002).

Cryptographic-Speech-Key Generation Architecture Improvements

L. Paola García-Perera, Juan A. Nolazco-Flores, and Carlos Mex-Perera

Computer Science Department, ITESM, Campus Monterrey
Av. Eugenio Garza Sada 2501 Sur, Col. Tecnológico
Monterrey, N.L., México, C.P. 64849
{paola.garcia,jnolazco,carlosmex}@itesm.mx

Abstract. In this work we show a performance improvement of our system by taking into account the weights of the mixture of Gaussians of the Hidden Markov Model. Furthermore and independently tunning of each of the phoneme Support Vector Machine (SVM) parameters is performed. In our system the user utters a pass phrase and the phoneme waveform segments are found using the Automatic Speech Recognition Technology. Given the speech model and the phoneme information in the segments, a set of features are created to train an SVM that could generate a cryptographic key. Applying our method to a set of 10, 20, and 30 speakers from the YOHO database, the results show a good improvement compared with our last configuration, improving the robustness in the generation of the cryptographic key.

1 Introduction

The generation of a cryptographic key based on biometrics, i.e. voice, face, fingerprints [13], is nowadays acquiring great importance because of security issues. The advantage of having a cryptographic key based on biometrics is that it simultanously act as a password for access control and as a key for encryption of data that will be stored or transmited. Moreover, given the biometric information it is also possible to generate a private key and a public key. Since in biometrics the characteristics are unique for each individual, the key generated will be difficult to guess. For that reason, having a key generated by a biometric is highly desirable.

From all the biometrics, voice was choosen in this research because a user can have the flexibility of changing a pass phrase when he requires it, or the system can also ask for a repetition of a random phrase, preventing unauthorised users access the system.

The results obtained in our previous work showed the potentiality of our system architecture [5–7]. Therefore, the purpose of this paper is to present the outcomes that improve our last results by considering the Gaussian weights in the interface between recognition and classification, and by performing a phoneme classification tunning. In this research the computer system consistently generates a cryptographic key based on the user's utterance and its matching pass

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 579–585, 2005.

phrase. In addition, a more flexible way to produce a key in which the exact control of the assignation of the key values is available.

The main challenge of this research is to find a method to produce a key with the characteristics already described. To achive good results we used speech processing and support vector machine techniques. Firstly, the speech signal is processed using an Automatic Speech Recogniser (ASR), from which a model and a phoneme based segmentation is obtained. Next, a feature generator handles the ASR output data to obtain suitable sets for the Support Vector Machine (SVM). Finally, the SVM classifies the users and the key is obtained. A general view of the system architecture is shown in Figure 1 and will be discussed in the following sections.

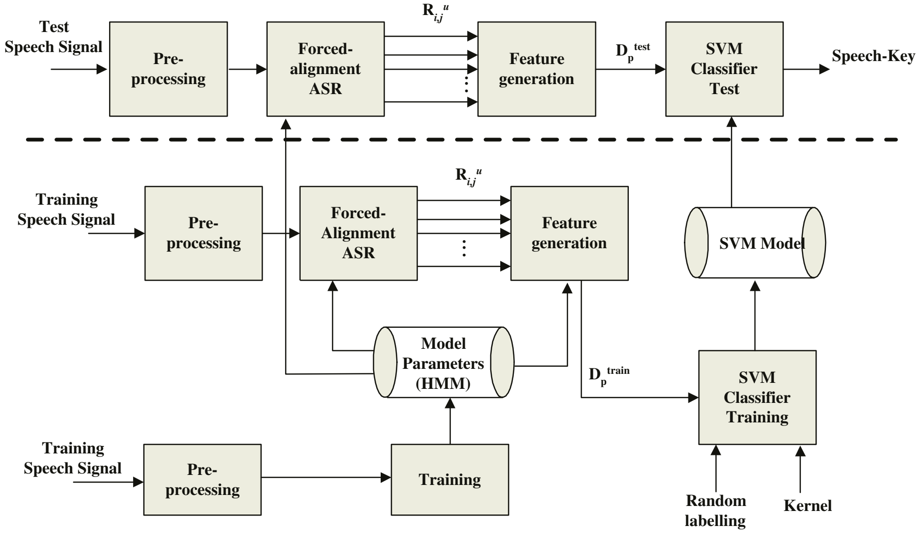


Fig. 1. System Architecture

2 Speech Processing

The primary task of this stage is to obtain the transcription and the starts and ends of the phonemes per user utterance. The speech signal is divided into short windows and the *Mel Frequency Cepstral Coefficients* (MFCC) are obtained. As a result a 13-dimension vector, 12-dimension MFCC followed by one energy coefficient is formed. To emphasize the dynamic features of the speech in time, the time-derivative (Δ) and the time-acceleration (Δ^2) of each parameter is calculated [11].

Afterwards, the ASR configured as a forced alignment recogniser provides a model and the starts and ends of the phonemes in a utterance. The ASR is based on a 3 state, left-right, Gaussian-based continuous Hidden Markov Model (HMM). Instead of words, the phonemes were selected because it is possible to generate larger keys with shorter length sentences. Assuming the phonemes are modelled with a three-state left-to-right HMM, and assuming the middle state is the most stable part of the phoneme representation, let,

$$C_i = \frac{1}{K} \sum_{l=1}^{K} W_l G_l, \tag{1}$$

where G is the mean of a Gaussian, K is the total number of Gaussians available in that state, W_l is the weight of the Gaussian and i is the index associated to each phoneme.

3 Phoneme Feature Generation

Given the phonemes' segments, the MFCCs for each phoneme in the utterances can be arranged forming the sets $R^u_{i,j}$, where i is the index associated to each phoneme, j is the j-th user, and u is an index that starts in zero and increments every time the user utters the phoneme i.

Then, the feature vector is defined as

$$\psi^u_{i,j} = \mu(R^u_{i,j}) - C_i$$

where $\mu(R^u_{i,j})$ is the mean vector of the data in the MFCC set $R^u_{i,j}$, and $C_i \in \mathcal{C}_P$ is known as the matching phoneme mean vector of the model. Let us denote the set of vectors,

$$D_p = \{\psi^u_{p,j} \mid \forall\ u, j\}$$

where p is a specific phoneme.

Afterwards, this set is divided in subsets: D^{tr}_p and D^{test}_p. 80% of the total D_p are elements of D^{tr}_p and the remaining 20% form D^{test}_p. Then, $D^{train}_p = \{[\psi^u_{p,j}, b_{p,j}] \mid \forall\ u, j\}$ where $b_{p,j} \in \{-1, 1\}$ is the key bit or class assigned to the phoneme p of the j-th user.

4 Support Vector Machine

The Support Vector Machine is a particular instance of the kernel machines derived by Vapnik and Chervonenkis [1, 3]. Although SVM has been used for several applications, it has also been employed in biometrics [9, 10]. The basic task of this algorithm is to perform the classification of the input data into one of two classes. Firstly, the following set of pairs are defined $\{x_i, y_i\}$; where $x_i \in \mathbb{R}^n$ are the training vectors and $y_i = \{-1, 1\}$ are the labels. The SVM learning algorithm finds an hyperplane (w, b) such that,

$$\min_{x_i, b, \xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \tag{2}$$

$$\text{subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i \tag{3}$$
$$\xi_i \geq 0$$

where ξ_i is a slack variable and C is a positive real constant known as a tradeoff parameter between error and margin. Equations 2 and 3 can be transformed into a dual problem that can be solved as a quadratic programming (QP) problem. To extend the linear method to a nonlinear technique, the input data is mapped into a higher dimensional space by a function named ϕ. However, exact specification of ϕ is not needed: instead, the expression known as kernel $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ is defined.

In this work, we explored the SVM using a radial basis function (RBF) kernel to classify sets of features. Those features are based on MFCC vectors and are to be transformed into sets of binary numbers (key bits) assigned randomly. The RBF kernel is denoted as

$$K(x_i, x_j) = e^{(-\gamma||x_i - x_j||^2)},$$

where $\gamma > 0$. The SVM uses also a decision criteria, which depends on C, a tradeoff parameter between error and margin.

Firstly, the training set for each phoneme (D_p^{train}) is formed by assigning a one-bit random label ($b_{p,j}$) to each user. Since a random generator of the values (-1 or 1) is used, the assignation is different for each user. The advantage of this random assignation is that the key entropy grows significantly. Afterwards, by employing a grid search the parameters C and γ are tuned to optimise the results. In our previous research, we developed some results perfoming a suboptimal tunnning using just a pair of C and γ for all cases. However, in this new approach we use a suboptimal tunning for each phoneme; *i.e* each phoneme will have its own C and γ. Finally, a testing stage is performed using D_p^{test}.

This research considers just binary classes. The final key could be obtained by concatenating the bits produced by each phoneme. For instance, if a user utters two phonemes: */F/* and */AH/*, the final key is $K = \{f(D_{/F/}), f(D_{/AH/})\}$, thus, the output is formed by two bits.

5 Experimental Methodology and Results

The YOHO database [2, 4] was used to perform the experiments. YOHO contains clean voice utterances of 138 speakers of different nationalities. It is a combination lock phrases (for instance, “Thirty-Two, Forty-One, Twenty-Five”) with 4 enrollment sessions per subject and 24 phrases per enrollment session; 10 verification sessions per subject and 4 phrases per verification session. Given 18768 sentences, 13248 sentences were used for training and 5520 sentences for testing. Next, the utterances are processed using the Hidden Markov Models Toolkit (HTK) by Cambridge University Engineering Department [8] configured as a forced-alignment automatic speech recogniser. The important results of the speech processing stage are the mean vectors of the phonemes C_i in Equation 1 given by the HMM and the phoneme starts and ends of the utterances. The phonemes used are: /AH/, /AX/, /AY/, /EH/, /ER/, /EY/, /F/, /IH/, /IY/,/K/, /N/, /R/, /S/, /T/, /TH/, /UW/, /V/, /W/. We have used 10, 20 and 30 users for our experiments and a mixture of 8 Gaussians to compare the cases.

The D_p sets are formed following the method described. It is important to note that the cardinality of each D_p set can be different since the number of equal phoneme utterances can vary from user to user. Next, subsets D_p^{train} and D_p^{test} are constructed. For training, the number of vectors picked per user, per phoneme for generating the model is the same. Each user has the same probability to produce the correct bit per phoneme. However, the number of testing vectors that each user provided can be different. For this work, the key bit assignation is arbitrary. Thus, the keys have liberty of assignation, therefore the keys entropy can be easily maximised if they are given in a random fashion with a uniform probability distribution.

SVMLight by Thorsten Joachims was used to implement the classifier [12]. The behaviour of the SVM is given in terms of the average classification accuracy on test data for a given number of users. The average classification accuracy is computed by the ratio

$$\eta = \frac{\text{matches on test data for all phonemes and users}}{\text{total number of vectors in test data}}. \tag{4}$$

In this work we perfomed two experiments:

1. The goal of the first experiment was to evaluate the impact of the Gaussian weights. Therefore, we compared the performance of the system with and without considering the weights of the Gaussians. Table 1 shows the results of these experiments for a system with a mixture of 8 Gaussians, and for 10, 20, and 30 users.

Table 1. Average % of η with and without Gaussian weights for different number of users

number of users	% of η without weight	% of η using weights
10	92.32	92.51
20	89.9	89.99
30	88.79	88.8426

2. The purpose of our second experiment is to evaluate the advantage obtained by independently performing the tunning of the SVM parameters for each of the phonemes. Table 2 shows the results of this experiment for a mixture of 8 Gaussian and 10 users.

6 Conclusion

In this research we proposed a method to eficiently generate a cryptographic key from voice. We used the techniques of the automatic speech recogniser and the support vector machines to achieve this purpose.

From the results we have found that the method to distinguish phonemes of specific users is quite good and provides good results for any key and user. The

Table 2. % of η for different phonemes, using phoneme tunning and 10 users

Phoneme	10user 8gauss weight	PHONE TUNNING
/AH/	92.8389	93.0936
/AO/	94.6542	94.8381
/AX/	94.7563	95.3859
/AY/	98.0601	98.2973
/EH/	94.0936	95.238
/ER/	96.376	96.416
/EY/	88.9621	89.0155
/F/	85.8751	85.9399
/IH/	93.6509	93.6531
/IY/	93.5343	94.2708
/K/	86.146	87.126
/N/	97.7116	97.9107
/R/	88.2419	89.9046
/S/	88.7694	89.3375
/T/	91.5536	92.0274
/TH/	86.4367	86.7832
/UW/	95.5974	95.7973
/V/	95.2885	95.4017
/W/	91.6403	92.398

increment of the number of Gaussians and the tunning by phoneme facilitate the classification and better results are obtained.

For further study some exploration on error correction algorithms should be considered. Besides, future studies on a M-ary key can be useful to increase the number of different keys available for each user given a fixed number of phonemes in the passphrase.

Acknowledgments

The authors would like to acknowledge the Cátedra de Seguridad, ITESM, Campus Monterrey and the CONACyT project CONACyT-2002-C01-41372 who partially supported this work.

References

1. Boser, B., I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 1992.
2. Campbell, J. P., Jr. Features and Measures for Speaker Recognition. Ph.D. Dissertation, Oklahoma State University, 1992.
3. Cortes, C. and V. Vapnik. Support-vector network. Machine Learning 20, 273–297, 1995.

4. Higgins, A., J. Porter and L. Bahler. YOHO Speaker Authentication Final Report. ITT Defense Communications Division, 1989.
5. Garcia-Perera L. P., C. Mex-Perera and J. A. Nolazco-Flores. Multi-speaker voice cryptographic key generation. Accepted for publication in the 3rd ACS/IEEE International Conference on Computer Systems and Applications - January 2005
6. Garcia-Perera, L. P., C. Mex-Perera, J.A. Nolazco-Flores, Criptographic-speech-key generation using the SVM technique over the lp-cepstra speech space, International School on Neural Nets, Lecture Notes on Computer Sciences (LNCS) , Springer-Verlag, Vietri, Italy, Sept., 2004. Accepted for publication.
7. Garcia-Perera L. P., C. Mex-Perera and J. A. Nolazco-Flores. SVM Applied to the Generation of Biometric Speech Key A. Sanfeliu et al. (Eds.): CIARP 2004, LNCS 3287, pp. 637-644, 2004. Springer-Verlag Berlin Heidelberg 2004
8. Young,S., P. Woodland HTK Hidden Markov Model Toolkit home page. http://htk.eng.cam.ac.uk/
9. E. Osuna, R. Freund, and F. Girosi. Support vector machines: Training and applications. Technical Report AIM-1602, MIT A.I. Lab., 1996.
10. E. Osuna, R. Freund, and F. Girosi, Training Support Vector Machines: An Application to Face Recognition, in IEEE Conference on Computer Vision and Pattern Recognition, pp. 130-136, 1997.
11. L.R. Rabiner and B.-H. Juang. Fundamentals of speech recognition. Prentice-Hall, New-Jersey, 1993.
12. T. Joachims, SVMLight: Support Vector Machine, SVM-Light Support Vector Machine http://svmlight.joachims.org/, University of Dortmund, November 1999.
13. U. Uludag, S. Pankanti, S. Prabhakar and A.K. Jain, Biometric cryptosystems: issues and challenges, Proceedings of the IEEE , Volume: 92 , Issue: 6 , June 2004.

Performance of a SCFG-Based Language Model with Training Data Sets of Increasing Size*

Joan Andreu Sánchez[1], José Miguel Benedí[1], and Diego Linares[2]

[1] Depto. Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 Valencia, Spain
{jandreu,jbenedi}@dsic.upv.es
[2] Pontificia Universidad Javeriana - Cali
Calle 18 No. 118-250 Av. Cañasgordas. Cali, Colombia
dlinares@dsic.upv.es

Abstract. In this paper, a hybrid language model which combines a word-based n-gram and a category-based Stochastic Context-Free Grammar (SCFG) is evaluated for training data sets of increasing size. Different estimation algorithms for learning SCFGs in General Format and in Chomsky Normal Form are considered. Experiments on the UPenn Treebank corpus are reported. These experiments have been carried out in terms of the test set perplexity and the word error rate in a speech recognition experiment.

1 Introduction

Language modeling is an important aspect to consider in the development of speech and text recognition systems. N-gram models are the most extensively used models for a wide range of domains [1]. A drawback of n-gram models is that they cannot characterize the long-term constraints of the sentences of the tasks. Stochastic Context-Free Grammars (SCFGs) efficiently model the long-term relations of the sentences. The two main obstacles to using these models in complex real tasks are the difficulties of learning SCFGs and of integrating SCFGs.

With regard to the learning of SCFGs, taking into account the existence of robust techniques for the automatic estimation of the probabilities of the SCFGs from samples [7, 8, 11, 14], in this work, we consider other possible approaches for the learning of SCFGs by means of a probabilistic estimation process [11].

When the SCFG is in Chomsky Normal Form (CNF), an initial exhaustive ergodic grammar is iteratively estimated by using the *inside-outside* algorithm or the Viterbi algorithm [3, 7, 10, 11]. When a treebank corpus is available, it is possible to directly obtain an initial SCFG in General Format (GF) from the syntactic structures that are present in the treebank corpus. Then these SCFGs in GF are estimated by using the Earley algorithm [8, 14].

With regard to the problem of SCFG integration in a recognition system, several proposals have attempted to solve this problem by combining a word n-gram model

* This work has been partially supported by the Spanish MCyT under contract (TIC2002/04103-C03-03) and by *Agencia Valenciana de Ciencia y Tecnología* under contract GRUPOS03/031.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 586–594, 2005.

and a structural model in order to take into account the syntactic structure of the language [5]. We proposed a general hybrid language model in [3] along the same line. This was defined as a linear combination of a word n-gram model, which was used to capture the local relation between words, and a stochastic grammatical model, which was used to represent the global relation between syntactic structures. In order to capture the long-term relations between syntactic structures and to solve the main problems derived from large-vocabulary complex tasks, we also proposed a stochastic grammatical model defined by a category-based SCFG together with a probabilistic model of word distribution into the categories.

Previous works have shown that the weight of the stochastic grammatical model defined in [3, 8] was less than expected and most of the information was conveyed in the n-gram model. This seemed reasonable, because there was enough data to adequately estimate the n-gram model. However, the performance of the hybrid language model has not been adequately studied when there is little training data. Taking this idea into consideration, in this work, we propose to study the performance of the hybrid language model with training data sets of increasing size.

In the following section, we briefly describe the hybrid language model and the estimation of the models. Then, we present experiments with the UPenn Treebank corpus. The experiments have been carried out in terms of the *test set perplexity* and the *word error rate*.

2 The Language Model

An important problem related to statistical language modeling is the computation of the expression $\Pr(w_k|w_1 \ldots w_{k-1})$. The n-gram language models are the most widely used for a wide range of domains [1]. The n-gram model reduces the history length to only $w_{k-n+1} \ldots w_{k-1}$. The n-grams are simple and robust models and adequately capture the local restrictions between words. Moreover, the way to estimate the parameters of the model and the way to integrate it in a speech and text recognition system are well known. However, they cannot characterize the long-term constraints of the sentences in these tasks.

Some works have proposed combining a word n-gram model and a structural model in order to take into account the syntactic structure of the language [5, 12]. Along the same line, in [3], we proposed a general hybrid language model defined as a linear combination of a word n-gram model, which is used to capture the local relations between words, and a word stochastic grammatical model M_s, which is used to represent the global relation between syntactic structures and which allows us to generalize the word n-gram model:

$$\Pr(w_k|w_1 \ldots w_{k-1}) = \alpha \Pr(w_k|w_{k-n+1} \ldots w_{k-1}) + (1-\alpha) \Pr_{M_s}(w_k|w_1 \ldots w_{k-1}), \quad (1)$$

where $0 \leq \alpha \leq 1$ is a weight factor that depends on the task.

The first term of expression (1) is the word probability of w_k given by the word n-gram model. The parameters of this model can be easily estimated, and the expression $\Pr(w_k|w_{k-n+1} \ldots w_{k-1})$ can be efficiently computed [1].

In order to capture long-term relations between syntactic structures and to solve the main problems derived from large-vocabulary complex tasks, we proposed a stochastic grammatical model M_s defined as a combination of two different stochastic models: a category-based SCFG (G_c) and a stochastic model of word distribution into categories (C_w). Thus, the second term of expression (1) can be written as:

$$\mathrm{Pr}_{G_c,C_w}(w_k|w_1\dots w_{k-1}). \tag{2}$$

In this proposal, there are still two important questions to consider: the definition and learning of G_c and C_w, and the computation of expression (2).

2.1 Learning of the Models

Here, we explain the estimation of the models. First, we introduce some notation. Then, we present the framework in which the estimation process is carried out. Finally, we describe how the parameters of G_c and C_w are estimated.

A *Context-Free Grammar* (CFG) G is a four-tuple (N,Σ,P,S), where N is a finite set of non-terminal symbols, Σ is a finite set of terminal symbols, P is a finite set of rules, and S is the initial symbol. A CFG is in Chomsky Normal Form (CNF) if the rules are of the form $A\to BC$ or $A\to a$ ($A,B,C\in N$ and $a\in\Sigma$). We say that the CFG is in General Format (GF) if no restriction is imposed on the format of the right side of the rules. A *Stochastic Context-Free Grammar* (SCFG) G_s is defined as a pair (G,q), where G is a CFG and $q:P\to]0,1]$ is a probability function of rule application such that $\forall A\in N$: $\sum_{\alpha\in(N\cup\Sigma)^+} q(A\to\alpha)=1$. We define the *probability* of the derivation d_x of the string x, $\mathrm{Pr}_{G_s}(x,d_x)$, as the product of the probability application function of all the rules used in the derivation d_x. We define the *probability* of the string x as: $\mathrm{Pr}_{G_s}(x)=\sum_{\forall d_x}\mathrm{Pr}_{G_s}(x,d_x)$.

Estimation Framework. In order to estimate the probabilities of a SCFG, it is necessary to define both an objective function to be optimized and a framework to carry out the optimization process. In this work, we have considered the framework of Growth Transformations [2] in order obtain the expression that allows us to optimize the objective function.

In reference to the function to be optimized, we will consider the likelihood of a sample which is defined as: $\mathrm{Pr}_{G_s}(\Omega)=\prod_{x\in\Omega}\mathrm{Pr}_{G_s}(x)$, where Ω is a multiset of strings.

Given an initial SCFG G_s and a finite training sample Ω, the iterative application of the following function can be used to modify the probabilities ($\forall(A\to\alpha)\in P$):

$$q'(A\to\alpha)=\frac{\sum_{x\in\Omega}\frac{1}{\mathrm{Pr}_{G_s}(x)}\sum_{\forall d_x}\mathrm{N}(A\to\alpha,d_x)\mathrm{Pr}_{G_s}(x,d_x)}{\sum_{x\in\Omega}\frac{1}{\mathrm{Pr}_{G_s}(x)}\sum_{\forall d_x}\mathrm{N}(A,d_x)\mathrm{Pr}_{G_s}(x,d_x)}\,. \tag{3}$$

The expression $\mathrm{N}(A\to\alpha,d_x)$ represents the number of times that the rule $A\to\alpha$ has been used in the derivation d_x, and $\mathrm{N}(A,d_x)$ is the number of times that the non-terminal A has been derived in d_x. This transformation optimizes the function $\mathrm{Pr}_{G_s}(\Omega)$.

Algorithms which are based on transformation (3) are gradient descendent algorithms and, therefore, the choice of the initial grammar is a fundamental aspect since it affects both the maximum achieved and the convergence process. Different methods have been proposed elsewhere in order to obtain the initial grammar.

Estimation of SCFG in CNF. When the grammar is in CNF, transformation (3) can be adequately formulated and it becomes the well-known Inside-Outside (IO) algorithm [7]. If a bracketed corpus is available, this algorithm can be adequately modified in order to take advantage of this information and we get the IOb algorithm [11]. If we use only the best derivation of each string, then transformation (3) becomes the Viterbi-Score (VS) algorithm [10].

The initial grammar for these estimation algorithms is typically constructed in a heuristic fashion by constructing all possible rules that can be composed from a given set of terminals symbol and a given set of non-terminal symbols [3, 7].

Estimation of SCFG in GF. When the grammar is in GF, transformation (3) can be adequately computed by using an Earley-based algorithm [8, 14] (the IOE algorithm). When a bracketed corpus is available, the algorithm can be adequately modified by using a similar function to the one described in [11], and we get the IOEb algorithm [8]. If we use only the best derivation of each string, then transformation (3) becomes the Viterbi-Score (VSE) algorithm [8].

In these algorithms, the initial grammar can be obtained from a treebank corpus [4, 8].

Estimation of the Parameters of C_w. We work with a tagged corpus, where each word of the sentence is labeled with part-of-speech tags (POStag). From now on, these POStags are referred to as word categories in C_w and are the terminal symbols of the SCFG in G_c. The parameters of the word-category distribution, $C_w = \Pr(w|c)$ are computed in terms of the number of times that the word w has been labeled with the POStag c. It is important to note that a word w can belong to different categories. In addition, it may happen that a word in a test set does not appear in the training set, and, therefore, its probability $\Pr(w|c)$ is not defined. We solve this problem by adding the term $\Pr(\texttt{UNK}|c)$ for all categories, where $\Pr(\texttt{UNK}|c)$ is the probability for unseen words of the test set.

2.2 Integration of the Model

The computation of probability (2) can be expressed as:

$$\Pr_{G_c,C_w}(w_k|w_1 \dots w_{k-1}) = \frac{\Pr_{G_c,C_w}(w_1 \dots w_k \dots)}{\Pr_{G_c,C_w}(w_1 \dots w_{k-1} \dots)},$$

where

$$\Pr_{G_c,C_w}(w_1 \dots w_k \dots) \ . \tag{4}$$

represents the probability of generating an initial substring given G_c and C_w.

Expression (4) can be easily computed by a simple modification [3] of the LRI algorithm [6] when the SCFG is in CNF, and by an adaptation [8] of the *forward* algorithm [14] when the SCFG is in GF.

3 Experiments with the UPenn Treebank Corpus

In this section, we describe the experiments which were carried out to test the language model proposed in the previous section for training sets of increasing size.

The corpus used in the experiments was the part of the Wall Street Journal that had been processed in the UPenn Treebank project [9]. It contains approximately one million words distributed in 25 directories. This corpus was automatically labeled, analyzed and manually checked as described in [9]. There are two kinds of labeling: a POStag labeling and a syntactic labeling. The size of the vocabulary is greater than 49,000 different words; the POStag vocabulary is composed of 45 labels; and the syntactic vocabulary is composed of 14 labels. The corpus was divided into sentences according to the bracketing. For the experiments, the corpus was divided into three sets: training (see Table 1), tuning (directories 21-22; 80,156 words) and test (directories 23-24; 89,537 words). Sentences labeled with POStags were used to learn the category-based SCFGs, and sentences labeled with both POStags and words were used to estimate the parameters of the hybrid language model.

First, we present the perplexity results for the described task. Second, we present word error rate results on a speech recognition experiment. In both experiments, the hybrid language model has been tested with training data sets of increasing size. Preliminary results of these experiments appeared in [3, 8].

3.1 Perplexity Results

We carried out the experiments taking into account the restrictions considered in other works [3, 5, 8, 12]. The restrictions that we considered were the following: all words that had the POStag CD (cardinal number [9]) were replaced by a special symbol which did not appear in the initial vocabulary; all capital letters were uncapitalized; the vocabulary was composed of the 10,000 most frequent words that appear in the training.

Baseline Model. We now describe the estimation of a 3-gram model to be used as both a baseline model and as a part of the hybrid language model. The parameters of a 3-gram model were estimated using the software tool described in [13]. Linear discounting was used as the smoothing technique with the default parameters in order to compare the obtained results with results reported in other works [8]. The out-of-vocabulary words were used in the computation of the perplexity, and back-off from context cues was excluded.

Hybrid Language Model. In this section, we describe the estimation of the stochastic grammatical model, M_s, and the experiments which were carried out with the hybrid language model.

The parameters of the word-category distribution (C_w) were computed from the POStags and the words of the training corpus. The unseen events of the test corpus were considered as the same word `UNK`, and we assigned a probability based on the classification of unknown words into categories in the tuning set. A small probability ϵ was assigned if no unseen event was associated to the category.

With regard to the estimation of the category-based SCFG (G_c) of the hybrid model, we first describe the estimation of SCFGs in CNF, and we then describe the estimation of SCFGs in GF.

For initial SCFGs in CNF, a heuristic initialization based on an exhaustive ergodic model was carried out. This initial grammar in CNF had the maximum number of rules that can be formed using 35 non-terminal symbols and 45 terminal symbols. In this way, the initial SCFG had 44,450 rules. Then, the parameters of this initial SCFG were estimated using several estimation algorithms: the VS algorithm and the IOb algorithm. Note that the IO algorithm was not used to estimate the SCFG in CNF because of the time that is necessary per iteration and the number of iterations that it needs to converge.

For SCFG in GF, given that the UPenn Treebank corpus was used, an initial grammar was obtained from the syntactic information which is present in the corpus. Probabilities were attached to the rules according to the frequency of each one in the training corpus. Then, this initial grammar was estimated using several estimation algorithms based on the Earley algorithm: the VSE algorithm, the IOE algorithm, and the IOEb algorithm.

Finally, once the parameters of the hybrid language model were estimated, we applied expression (1). In order to compute expression (4), we used:

- the modified version of the LRI algorithm [3] with SCFGs in CNF, which were estimated as we described above;
- the modified version of the *forward* algorithm described in [8], with SCFGs in GF, which were estimated as described above.

The tuning set was used to determine the best value of α for the hybrid model (2), that is, the weight factor.

In order to study the influence of the size of the training data set on the learning of the hybrid language model, we carried out the following experiment. All the parameters of the hybrid language model were estimated for different training sets of increasing size. The same restrictions which have been described above for estimating the parameters of the models (the category-based SCFG, the word distribution into categories and the n-gram model) were considered for each training set. In addition, a new baseline was computed for each training set. The tuning and test sets were the same for all cases. The results obtained can be seen in Table 1.

Table 1 shows that the test set perplexity with the n-gram models increased as the size of the training set decreased. The test set perplexity with the hybrid language model improved in all cases. It is important to note that both the percentage of improvement and the weight of the grammatical part increased as the size of the training set decreased. For SCFG in GF, the percentage of improvement was better than the percentage of improvement for SCFG in CNF. These results are very significant because they show that the proposed hybrid language model can be very useful when little training data is available.

3.2 Word Error Rate Results

Here, we describe preliminary speech recognition experiments which were carried out to evaluate the hybrid language model. Given that our hybrid language model is not integrated in a speech recognition system, we reproduced the experiments described in [8, 12] in order to compare our results with those reported in those works.

The experiment consisted of rescoring a list of n best hypotheses provided by a speech recognizer that used a different language model. In our case, the speech recognizer and the language model were the ones described in [5]. Then the list was reordered with the proposed language model. In order to avoid the influence of the language model

Table 1. Test set perplexity, percentage of improvement and value of α for the hybrid model for SCFGs estimated with different estimation algorithms and training data sets of increasing size (in number of words).

Directories	00-02	00-04	00-06	00-08	00-10	00-12	00-14	00-16	00-18	00-20
Training size	142,218	232,392	328,551	391,392	487,836	590,119	700,717	817,716	912,344	1,004,073
n-gram baseline	253.4	231.2	211.2	203.5	197.5	189.0	181.4	174.4	171.2	167.3
HLM-VS	224.6	209.7	194.7	188.6	184.0	175.6	169.8	163.7	161.4	157.2
% improv.	11.4	9.3	7.8	7.3	6.8	7.1	6.4	6.1	5.7	6.0
α	0.61	0.66	0.70	0.72	0.74	0.75	0.76	0.78	0.79	0.79
HLM-IOb	190.8	185.9	174.9	169.6	166.3	157.9	151.9	145.2	143.8	142.3
% improv.	24.7	19.6	17.2	16.6	15.8	16.5	16.3	16.7	16.0	15.0
α	0.45	0.53	0.54	0.59	0.61	0.62	0.62	0.62	0.62	0.65
HLM-VSE	185.8	178.2	167.5	163.7	159.9	154.1	149.6	145.0	142.8	140.4
% improv.	26.7	22.9	20.7	19.6	19.0	18.5	17.5	16.9	16.6	16.1
α	0.47	0.54	0.58	0.60	0.61	0.63	0.65	0.66	0.67	0.67
HLM-IOE	184.9	177.0	166.7	162.5	158.5	152.4	147.6	143.1	140.9	138.6
% improv.	27.0	23.4	21.1	20.2	19.8	19.4	18.6	18.0	17.7	17.2
α	0.46	0.53	0.57	0.58	0.61	0.63	0.64	0.66	0.66	0.66
HLM-IOEb	188.7	180.9	169.7	165.3	161.8	155.6	151.0	146.4	144.1	142.1
% improv.	25.5	21.8	19.7	18.8	18.1	17.7	16.8	16.1	15.8	(15.1)
α	0.47	0.54	0.58	0.60	0.62	0.63	0.65	0.66	0.67	0.67

of the speech recognizer, it is important to use a large value of n; however, for these experiments, this value was lower.

The experiments were carried out using the DARPA ’93 HUB1 test setup. This test consists of 213 utterances read from the *Wall Street Journal* with a total of 3,446 words. The corpus comes with a baseline trigram model using a 20,000-word open vocabulary and is trained on approximately 40 million words.

The 50 best hypotheses from each lattice were computed using Cyprian Chelba’s A* decoder, along with the acoustic and trigram scores. Unfortunately, in many cases, 50 different string hypotheses were not provided by the decoder [12]. An average of 22.9 hypotheses per utterance were rescored.

The hybrid language model was used to compute the probability of each word in the list of hypotheses. The probability obtained using the hybrid language model was combined with the acoustic score, and the results can be seen in Table 2 along with the results obtained for different language models. The word error rate without language model, that is, using only the acoustic scores was 16.8.

Table 2 shows that in all cases, the hybrid language model slightly improved the n-gram model. However, the good results in perplexity did not correspond with the WER result. While the percentage of improvement in perplexity increased as the training data size decreased, the WER result did not reflect this improvement. It should be pointed out that better WER results were obtained in [12]. However, it should be noted that the model proposed in [12] is more complex. Whereas our stochastic grammatical model is simple, and it is learned by means of well-known estimation algorithms.

4 Conclusions

We have studied the performance of a SCFG-based language model using different training set sizes. One model uses a SCFG in CNF and the other uses a SCFG in GF.

Table 2. Word error rate results for several models, using different training sizes, and the best language model weight.

Directories	00-02	00-04	00-06	00-08	00-10	00-12	00-14	00-16	00-18	00-20
n-gram baseline	16.8	16.7	16.8	16.5	16.8	16.7	16.7	16.7	16.7	16.6
LM weight	2	3.5	3.5	3	3	5	2.5	5.5	3	5
HLM-IOb	16.7	16.5	16.7	16.3	16.4	16.3	16.3	16.3	16.3	16.0
LM weight	2.2	2.3	2.4	5.1	4	5.7	5.2	5	5.1	6
HLM-IOE	16.7	16.6	16.8	16.4	16.5	16.4	16.3	16.2	16.4	16.2
LM weight	4.2	5.0	5.7	5.1	5.2	5.9	5.4	6.1	5.9	4

Both models were tested in an experiment on the UPenn Treebank corpus. Both models showed good perplexity results, and their percentage of improvement increased as the size of the training set decreased. The WER results were not as good as the perplexity results, and the performance seemed to increase as the size of the training set increased.

For future work, we propose to test the proposed hybrid language models in other real tasks.

Acknowledgements

The authors would like to thank Dr. Brian Roark for providing them with the n-best lists for the HUB1 test set.

References

1. L.R. Bahl, F. Jelinek, and R.L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190, 1983.
2. L.E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3:1–8, 1972.
3. J.M. Benedí and J.A. Sánchez. Estimation of stochastic context-free grammars and their use as language models. *Computer Speech and Language*, 2005. To appear.
4. E. Charniak. Tree-bank grammars. Technical report, Departament of Computer Science, Brown University, Providence, Rhode Island, January 1996.
5. C. Chelba and F. Jelinek. Structured language modeling. *Computer Speech and Language*, 14:283–332, 2000.
6. F. Jelinek and J.D. Lafferty. Computation of the probability of initial substring generation by stochastic context-free grammars. *Computational Linguistics*, 17(3):315–323, 1991.
7. K. Lari and S.J. Young. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 4:35–56, 1990.
8. D. Linares, J.M. Benedí, and J.A. Sánchez. A hybrid language model based on a combination of n-grams and stochastic context-free grammars. *ACM Trans. on Asian Language Information Processing*, 3(2):113–127, June 2004.
9. M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

10. H. Ney. Stochastic grammars and pattern recognition. In P. Laface and R. De Mori, editors, *Speech Recognition and Understanding. Recent Advances*, pages 319–344. Springer-Verlag, 1992.
11. F. Pereira and Y. Schabes. Inside-outside reestimation from partially bracketed corpora. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 128–135. University of Delaware, 1992.
12. B. Roark. Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27(2):249–276, 2001.
13. R. Rosenfeld. The CMU statistical language modeling toolkit and its use in the 1994 ARPA csr evaluation. In *ARPA Spoken Language Technology Workshop*, Austin, Texas, USA, 1995.
14. A. Stolcke. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21(2):165–200, 1995.

Speaker Dependent ASRs for Huastec and Western-Huastec Náhuatl Languages

Juan A. Nolazco-Flores, Luis R. Salgado-Garza, and Marco Peña-Díaz

Departamento de Ciencias Computacionales, ITESM, Campus Monterrey
Av. Eugenio Garza Sada 2501 Sur, Col. Tecnológico
Monterrey, N.L., México, C.P. 64849
{jnolazco,lsalgado,motilio}@itesm.mx

Abstract. The purpose of this work is to show the results obtained when the latest technological advances in the area of Automatic Speech Recognition (ASR) are applied to the Western-Huastec Náhuatl and Huastec languages. Western-Huastec Náhuatl and Huastec are not only native (indigenous) languages in México, but also minority languages, and people who speak these languages usually are analphabetic. A speech database was created by recording the voice of native speaker when reading a set of documents used for native bilingual primary school in the official mexican state education system. A pronunciation dictionary was created for each language. A continuous Hidden Markov Models (HMM) were used for acoustical modeling, and bigrams were used for language Modeling. A Viterbi decoder was used for recognition. The word error rate of this task is below 8.621% for Western-Huastec Náhuatl language and 10.154% for Huastec language.

1 Introduction

Language Technologies, such as ASR and Text-to-Speech Synthesis, is mature in many languages, such as English and Spanish [10] [11] [13]. This allows many computer applications to be developed in these languages, for example educative software. In the same way, Language Technology,when applied to prehispanic, can also be used to develop applications for these languages.

In México there are around 295 native american languages [1]. The total number of persons who speaks this american indian languages is around 8% of the total population in México, this means around 7,000,000 [1]. These 295 languages are grouped in families[1], and some families are grouped in stocks[2] Western-Huastec Náhuatl is in the Uzo-aztec stock and the Huastec Language is inside the Maya family[3],

[1] A family is a group of languages that easily can be shown to be genetically related when the basic evidence is examined [14].) (In México, there are six families, Aztecan, Corachol, Cahita, Tarahumaran, Tepiman, Tubar [1]).

[2] A stock is a group of language families that are genetically related to each other but, because of the time depth involved, the evidence is more difficult to assemble. In México, there are three stocks (Uzo-aztec, otomangue, okano) [14].

[3] The maya family is a language independent of any stock [1]).

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 595–602, 2005.

The southern Uzo-Aztec stock, which comprises around 49 languages, is spoken for around 1,750,000 persons [1]. The Aztecan family which comprises 28 náhuatl languages is one of the most important ones in this stock with around 1,600,000 speakers [1]. Western-Huastec Náhuatl variant is the most popular with 410,000 persons speaking the language, spoken in 1500 communities [1] in the Huastec region of San Luis Potosi, México, where Tamasunchale city is the center of this region [1].

The mayan family, which comprises around 31 languages, is spoken for 3, 381, 300 persons [15], is the most diversified and populous language family of Meso-America. Huastec languages is one of these languages with 101,000 speakers [15]. The Huastec language is separated in time for 2,500 years and physically by more than 1,000 miles from the nearest other Mayan language [15]. The Huastec is spoken in the Huastec region of San Luis Potosi, being Aquismón and Tancahuits de Santos the cities with more speakers [16]. It is also spoken in Veracruz, being Tantoyuca the city with more speakers [1] [16].

Since most of the persons who speaks these languages are analphabet every year the percentage of persons, compared with the total population of the region, who speaks this language is diminishing. Actually, from the 295 native american languages spoken in México 188 are endangered languages [2]. Speaking the majority language better equips children for success in the majority culture than speaking a less prestigious language [2].

However, preserving the language is important because the Language is the most efficient means of transmitting a culture, and it is the owners of that culture that lose the most when a language dies. Every culture has adapted to unique circumstances, and the language expresses those circumstances. Moreover, identity is closely associated with language [2]. The history tied up in a language will go unrecorded; the poetry and rhythm of a singular tongue will be silenced forever. The scientific search for Universal Grammar, the common starting point for all grammars that human children seem to be born with, depends on our knowing what all human languages have in common. The wholesale loss of languages that we face today will greatly restrict how much we can learn about human cognition, language, and language acquisition at a time when the achievements in these arenas have been greater than ever before [2].

In this work, we propose to develop an ASR systems for the Huastec and Western-Huastec Náhuatl language using continuous HMM and bigrams. We use this technology because continuous HMM is the most successful acoustic modeling technique and bigrams is also a very successful language modeling technique. Moreover, we believe that native people will be more interested in their own language, when, thanks to this and other studies, they knew that other people is interested in his their languages.

This paper is organized as follows. In section 2, the náhuatl language features are explained. In section 3, the huastec language features are explained. In section 4, the database features are explained. In section 5, the system architecture description is defined. In section 6, the experiments and results are given. Finally in section 7, the comments and conclusions are given.

2 Náhuatl Language

The Náhuatl is well know because it was the language of the Aztecs Empire of central México when spanish arrived. However, is less known that there are 28 types of Náhautl, some of them with less than 1000 speakers [1]. This work is concern with the Western-Huastec Náhuatl, which is the one with more speakers of the Uzo-Aztec languages.

Originally the Náhuatl language writings were a mixture of pictures of three classes: pictogram, ideograms and phonograms [5][4]. When Spanish arrived to mexican culture, one of their first task was to adapt the náhuatl language to the spanish alphabet. Therefore, now the bilingual education in Mexico is with alphabet writing.

Náhuatl language is highly agglomerative, which means that words are formed by a root and a high number of prefixes and suffixes [5]. Therefore, the words in this kind of languages includes a lot of information and potentially each word can be very long and the number of words in the language is very high. As an example, the following written in English [6]:

> "This book is for indigenous boys and girls who are studying basic school with the aim to help them how to read and write the indigenous language spoken in its community"

will look as follows in Western-Huastec Náhuatl language:

> "Ni amochtli tijtlaliaj inmako ockichpilmej uan siuapilmej ankij momachtiaj Tlen eyi uan tlen naui xiuitl tlen se ixelka tlamachtilistli, pampa moneki kiyekosej tlaixpouasej uan tlajkuilosej ika inineltlajtol tlen ika tlajtouaj ipan inchinanko"

Table 1 shows the Western-Huastec Náhuatl language's phonemes used in this work [5]. There are some important pronunciation rules. First, the j not pronounced when it is at the end of a sentences, and some speaker ignore it even it is inside of the sentence. The rest of the letter are pronounced as they are written, with the following exceptions: when letter C is before letters E and I, then it is pronounced as the phoneme /S/.When the letters H followed by U is located before A, E and I, then it is pronounced as /W/. Given the text the pronunciation rules and the list of phonemes, and the labels in the speech database and the phonemes we create a pronunciation dictionary.

3 Huastec Language

This language is also known as Tenek language. Originally the huastec language writings were a mixture of pictures of three classes: pictogram, ideograms and

[4] In a pictogram an object is represented with one picture, in a ideogram something or an idea is represented with a picture, phonogramas a ?syllable? or phone represented with a picture.

Table 1. Phonemes used in this work for Western-Huastec-Náhuatl and Huastec languages.

Manner	WH Náhuatl	Huastec	Example
Vowels	a	a	h**o**od
	e	e	h**ea**d
	i	i	h**ee**d
	o	o	h**oe**d
	u	u	h**oo**d
Plosives	b		**b**oot
	p	p	**p**ea
	t	t	**t**ea
	k	k	**k**ick
Fricatives	s	s	**s**o
	S	S	**s**how
Nasals	m	m	**m**om
	n	n	**n**oon
Semivowels Glides	w	w	**w**ant
	y	y	**y**ard
Semivowels Liquids	l	l	l
Affricatives	C	C	**ch**urch (written with letter x.)
Others	tl		t and l pronounces as one sound
	tz		t and z pronounces as one sound
		dh	d and h pronounces as one sound

phonograms [15]. It is believed that the first book written in a language different to Náhuatl was in Huastec, and it was called ”Doctrina Cristina en Lengua Guasteca” [16].

Huastec language is highly agglomerative, which means that words are formed by a root and a high number of prefixes and suffixes [16]. Therefore, the words in this kind of languages includes a lot of information and potentially they can be very large, and the number of words in the language is very high. As an example, the following written in English [6]:

> “This book is for indigenous boys and girls who are studying basic school with the aim to help them how to read and write the indigenous language spoken in its community”

will look as follows in huastec language:

> “Axé xi dhuchadh úw, jats abal ka pidhanchik an ts'ik'ách ani an kwitól axi k'wátchik ti exóbal ti al an k'a'aál pejach tin k'a'ál exobintal; axé, jats abal kin ne'ets with'a'chik ti dhuchum ani ti ajum tin tének káwintal.”

Table 1 shows the huastec language's phonemes used in this work [5]. There are some important pronunciation rules. First, the *j* is not pronounced when it is at the end of a sentences, and some speaker ignore it even it is inside of the sentence. The rest of the letter are pronounces as they are written, with the following exception: when letter *dh* is read is pronounced as it where one sound.

4 Databases

In order to facilitate our labeling process, for our databases recording we selected some text books used for native language bilingual education in México [6]. Moreover, this was also very convenient to facilitate the labeling process.

In our speech database construction we ask to person to read lessons from the selected textbooks [6]. The number of different words in Western-Huastec Náhuatl are 759,this database contains around 1 hour of recorded data from two speakers, a man and a woman. The number of different words in Huastec are 319,this database contains around 1 hour of recorded data from two speakers, a man and a woman. All the participants with ages between 20 and 25 years old. In both cases, the speech waveform was sampled at 16,000 KHz.

5 System Architecture Description

The CMU SPHINX-III systems is a HMM-based speech recognition system capable of handling large vocabulary. The architecture of this system is shown in Figure 1. As can be observed in this figure the analog signal is sampled, and converted to MFCC coefficients, then the MFCC's first and second derivatives are concatenated [8], i.e. if the number of MFCC is 13 then the total dimension of the feature vector would be 39.

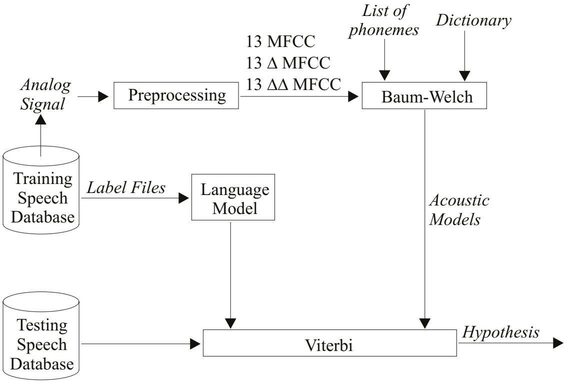


Fig. 1. CMU SPHINX-III ASR architecture.

The acoustic models is also obtained using the SPHINX-III tools. This tools use a Baum-Welch algorithm to train this acoustic models [12]. The Baum-Welch algorithm needs the name of the word units to train as well at the label and feature vectors. The SPHINX-III system allows us to modelate either discrete, semi continuous or continuous acoustic models. In SPHINX-III, system tools allow to select as acoustic model either a phone set, a triphones set or a word set.

The language models are obtained using the CMU-Cambridge statistical language model toolkit version 2.0 [9]. The LM aim is to reduce the perplexity of the task, by predicting the following word based in the words' history. N-grams is the easiest technique with very good results. If all the n-grams are not contained in the language corpus, smoothing techniques need to be applied. In the CMU-Cambridge language model toolkit, unigram, bigrams or trigrams can be configured for this tool, as well as four types of discount model: Good Turing, Absolute, Linear and Witten-Bell.

Using an acoustical model and a language model a Viterbi decoder obtains the best hypothesised text.

6 Experiments

The configuration of the SPHINX-III system is described. Thirteen mel-frequency cepstral coefficients (mfcc) were used. First and Second derivatives were calculated, therefore the feature vector was 39 elements. The speech lower frequency was 300 Hz and the speech higher frequency was 7,000 Hz. The frame rate was set to 50 frames per second. A 30ms Hamming window was used. A 512 samples FFT length was used. The number of filterbanks was set to 40. Five states continuous HMM were used as acoustic modeling technique and bigrams was used as a language modeling technique. Simple phones were used as the word unit. Since our corpus is a small corpus and the number of words is very large, we develop experiments using different smoothing techniques. Table 2 shows the experimental results for Western-Huastec Náhuatl language and Table 3 shows the experimental results for Huastec language. As expected the Witten-Bell discount strategy was the one with better results.

Table 2. WER results for Western-Huastec Náhuatl language over several Gaussian distributions and language model configurations.

Number of	Language Model discounting strategy			
Gaussians	**Good-Turing**	**Linear**	**Absolute**	**Witten-Bell**
4	6.80%	8.43%	6.80%	4.50%
8	6.71%	8.62%	6.80%	4.31%
16	6.80%	8.53%	6.80%	4.22%
32	6.90%	8.53%	6.80%	4.12%
64	6.90%	8.53%	6.80%	4.02%

Table 3. WER results for Huastec Language and over several Gaussian distributions and language model configurations.

Number of	Language Model discounting strategy			
ans	**Good-Turing**	**Linear**	**Absolute**	**Witten-Bell**
4	9.13%	9.83%	8.23%	6.94%
8	9.13%	10.03%	8.48%	6.56%
16	8.74%	9.90%	7.97%	6.81%
32	8.87%	10.15%	8.36%	6.68%
64	8.87%	10.15%	8.36%	6.94%

7 Conclusions

In this work, we present the development of a prehispanic database for Western-Huastec Náhuatl and Huastec languages. We also show the results obtained when Automatic Speech Recognition technology is applied to these languages. Since people that speak prehispanic language do not usually read, then the main problem to develop speech models for prehispanic languages is the difficult to find people that read its own language.

We think that Speech Technology can be a catalizer in the effort to preserve the minority languages. Therefore, as a future work, in first place the database will be extended to include a larger number of speakers, the recording time will also be extended. Other languages technologies, such as Text-to-Speech technology is also planned to be applied. The goal is to better understand the language to develop educative software in these languages.

We also have to refine the phoneme list and the pronunciation rules. We are also planning to create databases for other minority languages, such as Mixteco and Cora.

Acknowledgements

The authors would like to acknowledge the head of the Secretaría de Educación del Estado de Nuevo León, Maestra María Yolanda Blanco García; to the head of Departamento de Educación Indígena, Lic. Luz Divina Tijerina Luna. The authors also would like to acknowledge CONACyT, project CONACyT-2002-C01-41372, who partially supported this work.

References

1. http://www.ethnologue.com.
2. http://yourdictionary.com/elr/living.html.
3. Constitución Política de los Estados Unidos Mexicanos.
4. Plan y Programa de Estudio para la Educación Primaria, SEP, México, 1993.
5. Sullivan, T.O., Compendio de la Gramática Náhuatl, Ejercicios, UNAM, Instituto de Investigaciones Históricas, Second Edition, 1992.
6. Canales Juarez, G., Mendez González, R., Hernández Miranda, J., Roque Cerroblanco, E., "*Nauatlajtoli tlen uaxtekapaj tlali, Lengua náhuatl, Region Huasteca, Hidalgo*", Third and fourth grade, SEP, 1993.
7. http://www.ldc.upenn.edu.
8. Deller, J.R., Proakis, J.G., Hansen, J.H.L., Discrete-Time Processing of Speech Signals, Prentice Hall, Sec. 6.2, 1993.
9. Clarkson, P., Rosenfeld, R., "Statistical Language Modelling using the CMU-Cambridge Toolkit", Proceedings of Eurospeech, Rodhes, Greece, 1997, 2707-2710.
10. Varela, A., Cuayáhuitl, H., Nolazco-Flores, J.A., "Creating a Mexican Spanish Version of the CMU SPHINX-III Speech Recognition System", CIARP, Springer Verlag, LNCS 2905:251-58.

11. Salgado-Garza, L.R., Stern, R., Nolazco, J.A.,"N-Best List Rescoring using Syntactic Trigrams", MICAI 2004: Advances in Artiticial Intelligence LNAI 2972, Springer Verlag, 2004, LNAI 2972:79-88.
12. Dempster, A.P., Laird, N.M., Rubin, D.B., "Maximum likelehood for incomplete data via the EM algorithm", J. Roy. Stat. Soc., Vol. 39, No. 1, 1-38, 1977.
13. Huerta, J.M., Chen, S., Stern, R.M.: "The 1998 CMU SPHINX-3 Broadcast News Transcription System", Darpa Broadcast News Workshop, 1999.
14. Dryer, M. S., Large Linguistic areas and lang. samp. *Studies in Language*, 13:257-92, 1996.
15. "Meso-American Indian Languages". Encyclopedia Britannica. 2004. Encyclopedia Britannica Online. 14 May 2004
http://0-search.eb.com.millenium.itesm.mx:80/eb/article?eu=118158.
16. Grossner-Lerner, E., Los tenek de San Luis Potosi, INAH, 1991.

Part IX

Natural Language Analysis

Phrase-Based Alignment Models for Statistical Machine Translation

Jesús Tomás[1], Jaime Lloret[2], and Francisco Casacuberta[1]

[1] Instituto Tecnológico de Informática
Universidad Politécnica de Valencia, 46071 Valencia, Spain
{jtomas,jlloret,fcn}@upv.es
[2] Departamento de Comunicaciones
Universidad Politécnica de Valencia, 46071 Valencia, Spain

Abstract. The first pattern recognition approaches to machine translation were based on single-word models. However, these models present an important deficiency; they do not take contextual information into account for the translation decision. The phrase-based approach consists in translating a multiword source sequence into a multiword target sequence, instead of a single source word into a single target word. We present different methods to train the parameters of this kind of model. In the evaluation phase of this approach, we obtained interesting results in comparison with other statistical models.

1 Introduction

Statistical machine translation has been formalized in [1]. This approach defines a translation model by introducing the concept of alignment, which defines the correspondence between the words of source sentences and target sentences. The optimal parameters of this model are estimated using training sets of pairs of sentence. The most common statistical translation models can be classified as *single-word based* (SWB) alignment models. Models of this kind assume than a source word is generated by only one target word [1][2]. This assumption does not correspond to the nature of natural language; in some cases, we need to know a multiword sequence in order to obtain a correct translation.

Recent works present a simple alternative to these models, the *phrase-based* (PB) approach the probability of a multiword sequence in a source sentence being translated to another multiword sequence of words in the target sentence. These bigger units allow us to represent contextual information in an explicit and easy way. One shortcoming of the PB alignment models is the generalization capability. If a sequence of words has not been seen in training, the model cannot reorder it properly.

The organization of the paper is as follows. First, we review the statistical approach to machine translation. Second, we propose a *monotone phrase translation* model and propose different methods to estimate the parameters. Then, we propose a *non-monotone phrase translation* model. Finally, we report some experimental results. The system was tested using a corpus in several languages.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 605–613, 2005.

2 Statistical Translation

The goal of statistical machine translation is to translate a given source language sentence $\mathbf{s} = \mathbf{s}_1 ... \mathbf{s}_J \equiv \mathbf{s}_1^J$ to a target sentence $\mathbf{t} = \mathbf{t}_1 ... \mathbf{t}_I \equiv \mathbf{t}_1^I$, where J is the number of words in the source sentence and I is the number of words in the target sentence. The methodology used with stochastic translation [1] is based on the definition of a function $\Pr(\mathbf{t}|\mathbf{s})$ that returns the probability of translating a given source sentence, $\mathbf{s}$, into a target sentence, $\mathbf{t}$. Once this function is estimated, the problem can be reduced to computing a sentence $\mathbf{t}$ that maximizes the probability $\Pr(\mathbf{t}|\mathbf{s})$ for a given $\mathbf{s}$. Using Bayes' theorem, the maximization is reduced to:

$$\hat{\mathbf{t}} = \arg\max_{\mathbf{t}} \Pr(\mathbf{t}) \cdot \Pr(\mathbf{s}|\mathbf{t}) \tag{1}$$

Equation 1 summarizes the following three matters to be solved: A *target language model* ($P(\mathbf{t})$) is needed to distinguish valid sentences from invalid sentences in the target language, a *translation model* ($P(\mathbf{s}|\mathbf{t})$) and the design of *an algorithm to search* for the sentence $\mathbf{t}$ that maximizes this product.

In practice, instead of using equation 1 with the above models, the following heuristic decision rule can be used [3, 4]:

$$\hat{\mathbf{t}} \approx \arg\max_{\mathbf{t}} P(\mathbf{t}) \cdot P(\mathbf{t}|\mathbf{s}) \tag{2}$$

This equation uses a direct model for the translation model $P(\mathbf{t}|\mathbf{s})$; that is, we estimate the probability of the target sentence $\mathbf{t}$, given the source sentence $\mathbf{s}$. Some advantages of using this decision rule is that the search procedure can be performed more efficiently [4] and it is easier to include additional dependencies into the models [3].

3 Phrase-Based Alignment Models

The principal innovation of the PB translation alignment model [5][6] is that it attempts to calculate the translation probabilities of multiword sequences rather than of only single words. Figure 1 shows the same sentence written in five different languages.

Se requerirá	una acción	de la Comunidad	para la	total puesta en práctica
É necessária	uma acção	por parte da Comunidade	para pôr	plenamente em prática
Sarà necessaria	un'azione	della Comunità	per dare	piena attuazione
Une action	est nécessaire	au niveau communautaire	afin de	mettre pleinementen oeuvre
Action	is required	by the Community	in order to	implement fully

Fig. 1. Equivalent multiword sequences in a sentence in Spanish, Portuguese, Italian, French and English.

As can be seen in figure 1, we join words that are translated together in a natural way. The alignment between pairs of phrase sequences can be monotone-constrained. In the example, the first three sentences are monotone-translated.

In our model, only phrases of contiguous words are assumed and there are the same number of source phrases as target phrases (say K phrases). Introducing the size of target phrases through a function μ:

$$\mu : \{1, \ldots, K\} \rightarrow \{1, \ldots, I\} : \mu_k \geq \mu_{k-1} \quad 1 < k \leq K \quad \& \quad \mu_K = I \quad (\mu_0 = 0) \ ,$$

and, introducing the size of source phrases through a function γ:

$$\gamma : \{1, \ldots, K\} \rightarrow I\!N^+ : \gamma_k \geq \gamma_{k-1} \quad 1 < k \leq K \quad (\gamma_0 = 0) \ .$$

Let $J = \gamma_K$, the length of the source sentence. Then,

$$\Pr(\mathbf{s}|\ \mathbf{t}) = \sum_K \sum_{\mu_1^K} \sum_{\gamma_1^K} \Pr(K, \mu_1^K, \gamma_1^K |\ \mathbf{t}_1^I) \cdot \Pr(\mathbf{s}_1^J |\ \mathbf{t}_1^I, K, \mu_1^K, \gamma_1^K) \ . \qquad (3)$$

In translation, the corresponding K target phrases can be in a different order given by a permutation:

$$\alpha : \{1, \ldots, K\} \rightarrow \{1, \ldots, K\} : \alpha(k) = \alpha(k') \ \text{ iff } \ k = k' \ .$$

Introducing the permutation function in equation 3 and by factorizing some of the factors,

$$\Pr(\mathbf{s}|\mathbf{t}) = \sum_K \sum_{\mu_1^K} \sum_{\gamma_1^K} \Pr(K, \mu_1^K, \gamma_1^K |\ \mathbf{t}_1^I) \cdot \sum_{\alpha_1^K} \prod_{k=1}^{K} \Big(\Pr(\alpha_k | \mathbf{t}_1^I, K, \mu_1^K, \gamma_1^K, \alpha_1^{k-1}) \cdot \Pr(\mathbf{s}_{\gamma_{\alpha_k - 1}+1}^{\gamma_{\alpha_k}} | \mathbf{t}_1^I, K, \mu_1^K, \gamma_1^K, \alpha_1^k, \mathbf{s}_1^{\gamma_{\alpha_1}}, \ldots, \mathbf{s}_{\gamma_{\alpha_{k-1}-1}+1}^{\gamma_{\alpha_{k-1}}}) \Big) \qquad (4)$$

3.1 Monotone Phrase-Based Alignment Models

Different approaches can be adopted for equation 4. The simplest one can assume that all source and target segmentations have the same probability $(\Pr(K, \mu_1^K, \gamma_1^K |\ \mathbf{t}_1^I) = p_I)$. We can also assume that each source phrase depends only on the target phrase that has been aligned. And, if monotonicity is assumed $(\alpha_k = k)$, then, the source phrase in position k depends only on the target phrase in position k.

$$\Pr(\mathbf{s}|\mathbf{t}) \approx P(\mathbf{s}|\mathbf{t}) = p_I \cdot \sum_K \sum_{\mu_1^K} \sum_{\gamma_1^K} \prod_{k=1}^{K} p(\mathbf{s}_{\gamma_{k-1}+1}^{\gamma_k} | \mathbf{t}_{\mu_{k-1}+1}^{\mu_k}) \ . \qquad (5)$$

The parameter p_I is not relevant for translation and will be omitted. Thus, the only parameters of this model are $p(\widetilde{s}|\widetilde{t})$, that estimate the probability that the word group, $\widetilde{t}$, is translated to the word group $\widetilde{s}$.

3.2 Learning Monotone Phrase-Based Alignment Models

Training with a Sentence-Aligned Corpus. Given a sentence-aligned corpus $\mathcal{T}$, composed by a sample of pairs of sentences $(\mathbf{s}, \mathbf{t})$, the maximum likelihood criterium tries to estimate the parameters $p(\widetilde{s}|\widetilde{t})$ that maximize $\prod_{(\mathrm{s},\mathrm{t})\in\mathcal{T}} P(\mathbf{s}|\mathbf{t})$ subject to the constraints that hold for each $\widetilde{t}$: $\sum_{\widetilde{s}} p(\widetilde{s}|\widetilde{t}) = 1$. The corresponding reestimation formula is [5]:

$$p(\widetilde{s}|\widetilde{t}) = \lambda_{\tilde{t}} \cdot \sum_{(\mathrm{s},\mathrm{t})\in\mathcal{T}} p_I \cdot \sum_{K} \sum_{\mu_1^K} \sum_{\gamma_1^K} \left(\prod_{k=1}^{K} p(\mathbf{s}_{\gamma_{k-1}+1}^{\gamma_k} | \mathbf{t}_{\mu_{k-1}+1}^{\mu_k}) \cdot \right.$$
$$\left. \sum_{l=1}^{K} \delta(\widetilde{s} = \mathbf{s}_{\gamma_{l-1}+1}^{\gamma_l}) \cdot \delta(\widetilde{t} = \mathbf{t}_{\mu_{l-1}+1}^{\mu_l}) \right) \qquad (6)$$

where $\lambda_{\tilde{t}}$ is a normalization factor and δ is defined as: $\delta(true) = 1$, $\delta(false) = 0$.

Training with a Word-Aligned Corpus. The parameters of the model can also be obtained using a word-aligned corpus [6, 7]. These alignments can be obtained from SWB models [1] using the public available software GIZA++ [8].

The method for dealing with phrases consists of two steps. In the first step, a set of bilingual phrases from the word aligned corpus is obtained. In the second step, the parameters of the PB model are estimated.

Extracting Bilingual Phrases. Basically, a bilingual phrase consists of a pair of m consecutive source words that has been aligned with n consecutive target words. Different criteria can define the set of bilingual phrases BP in the sentence pair $(\mathbf{s}; \mathbf{t})$ with an alignment $\mathbf{a}$. Three different criteria are tried that are illustrated in figure 2.

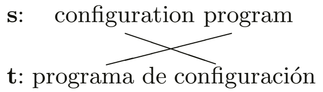
The BP_1 criterion considers $\mathbf{s}_{j_1}^{j_2} - \mathbf{t}_{i_1}^{i_2}$ as a bilingual phrase if all the words in $\mathbf{s}_{j_1}^{j_2}$ are aligned with a word in $\mathbf{t}_{i_1}^{i_2}$, and vice versa [6]. The BP_2 criterion is similar to the previous one, but, it allows some words in $\mathbf{s}_{j_1}^{j_2}$ or in $\mathbf{t}_{i_1}^{i_2}$ to be unaligned [9]. The BP_3 criterion forces the bilingual phrases to be extracted in a monotone way [10]. That is, it does not permit the extraction of a bilingual phrase if there is a word at the left of the source phrase that has been aligned to a word at the right of the target phrase (or vice versa).

The bilingual phrase extraction can be performed using one translation direction (s→t) or using symmetrization [4]. The last approach consists in obtaining two word-aligned corpora, each one in one direction (s→t and t→s). Then, we can extract the phrases in both corpora using one of the above BP proposals.

Estimating the Parameters. The estimation of the parameters of the model can be done via relative frequencies, for each pair of segments $(\mathbf{s}, \mathbf{t})$[6]:

$$p(\widetilde{s}|\widetilde{t}) = \frac{N(\widetilde{s}, \widetilde{t})}{N(\widetilde{t})} \qquad (7)$$

where $N(\widetilde{t})$ denotes the number of times that phrase $\widetilde{t}$ has appeared, and $N(\widetilde{s}, \widetilde{t})$ is the number of times that the bilingual phrase $(\widetilde{s}, \widetilde{t})$ has appeared.



BP_1={configuration-configuración, program-programa}
BP_2={configuration-configuración, program-programa,configuration-de configuración, program-programa de, configuration program-programa de configuración}
BP_3={configuration program-programa de configuración}

Fig. 2. Example of extraction a set of bilingual phrases from a word aligned sentence using three different criteria.

Another way of estimating the parameters can be carried out: to maximize $\prod_{(\mathrm{s},\mathrm{t})\in\mathcal{T}} P(\mathbf{s}|\mathbf{t})$ subject to the constraint that the PB-alignment in this function be consistent with the word-alignment in the training [9]. Using standard maximization techniques we obtain an equation similar to equation 6, but in this case the sum on l is extended to those terms such that $(\widetilde{s},\widetilde{t}) \in BP_i (i \in \{1,2,3\})$.

3.3 Non-monotone Phrase-Based Models

In this case, reordering of phrase sequences is permitted and the probability of an alignment α_k can be depend on the last alignment α_{k-1} (first-order alignment). Then, equation 4 becomes:

$$P(\mathbf{s}|\mathbf{t}) = p_I \cdot \sum_{K} \sum_{\mu_1^K} \sum_{\gamma_1^K} \sum_{\alpha_1^K} \prod_{k=1}^{K} p(\alpha_k|\ \alpha_{k-1}) \cdot p(\mathbf{s}_{\gamma_{\alpha_k-1}+1}^{\gamma_{\alpha_k}} | \mathbf{t}_{\mu_{k-1}+1}^{\mu_k}) \ , \qquad (8)$$

For the distortion model, we assume an alignment that depends only on the distance of the two phases [4]:

$$p(\alpha_k|\alpha_{k-1}) = p_0^{|\gamma_{\alpha_k} - \gamma_{\alpha_{k-1}}|} \qquad (9)$$

4 Search Algorithm

The aim of the search in MT is to look for a target sentence $\mathbf{t}$ that maximizes the product $P(\mathbf{t}) \cdot P(\mathbf{t}|\mathbf{s})$. The generative process, which allows for the translation of a sentence in the monotone model, can be broken down into the following steps: First, the source sentence, $\mathbf{s}$, is segmented into K phrases, $\tilde{s}_1^K$. Then, each phrase, $\tilde{s}_k$, is translated to the corresponding target phrase, $\tilde{t}_k$. The target sentence is built by concatenating the target phrases in the same order as in the source phrases, $\mathbf{t} = \tilde{t}_1^K$.

Our search is based on the multi-stack-decoding algorithm [11]. The basic multi-stack-decoding algorithm searches for only the best alignment, between the sentences $\mathbf{s}$ and $\mathbf{t}$ (We will refer to this method as Best alignment). However, in (4), we define the probability of translating $\mathbf{s}$ to $\mathbf{t}$ as the sum of all possible alignments. To solve this deficiency, we act as follows: We compute the real probabilities of all the complete hypotheses obtained using equation (5), and we take the largest one (We will refer to it as Add all alignments). We propose extending the search to allow for non-monotone translation. In this extension, several reorderings in the target sequence of phrases are scored with a corresponding probability [12].

5 Experimental Results

In order to evaluate the performance of these approaches, we carried out several experiments using the XRCE task. This corpus was compiled using some Xerox technical manuals published in several languages. This is a reduced-domain task that has been defined in the TransType2 project [13]. Table 1 presents some statistical information about this corpus after the pre-processing phase.

Table 1. XRCE corpus statistics. (K$\equiv \times 1,000$).

		English	Spanish	English	German	English	French
Training	Sentences pairs	55,761		49,376		52,844	
	Running words	665K	753K	633K	696K	587K	534K
	Vocabulary	7,957	11,051	7,790	9,922	7,664	19,297
Test	Sentences pairs	1,125		984		996	
	Running words	8K	10K	11K	12K	12K	12K
	Trigram PP	48	33	51	87	73	52

We selected the trigram model for the language model. For evaluation criterium we use *word error rate* (WER), the minimum number of substitution, insertion and deletion operations needed to convert the word string hypothesized by the MT into a given reference word string [4]. In the experiments the default configuration is: monotone direct model training with symmetric BP_2 and relative frequencies and maximum phrase length of 14 words.

In the formal description of the model, we do not limit the number of words in a phrase. However, in a practical implementation, we limit this parameter. Figure 3 shows the effect of this limitation on the translation results.

Table 2 shows the effect of the extraction criterion of the bilingual phrases. The best results were obtained with the BP_2 criteria using symmetrization.

Table 3 shows the results obtained using different types of training. The best results were obtained with a direct approach using relative frequencies. Table 4 compares the monotone and non-monotone search. For this task the results were similar.

In the TransType2 project several statistical translation models were tested. Table 5 compares the results obtained by GIATI [10] and the Phrase-Based model. The Phrase-Based model obtained the best results. In [14] some results have been presented using *alignment templates* [4]. With this approach, the WER was 28.9% for Spanish-English direction. The three methods have some

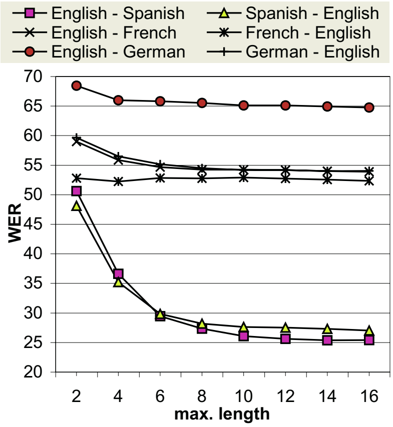


Fig. 3. Effect of maximum number of words in a phrase in WER.

Table 2. Effect of the extraction criterion of the set of bilingual phrases in the number of parameters and the WER. (s→t ≡ using one translation direction, s↔t ≡ using symmetrization, M ≡ ×1, 000, 000).

		English		Spanish		English		French		English		German	
		Spanish		English		French		English		German		English	
		s→t	s↔t	s→t	s↔t	s→t	s↔t	s→t	s↔t	s→t	s↔t	s→t	s↔t
BP_1	WER	45.2	26.1	27.5	26.9	55.8	52.2	51.8	50.8	64.9	63.8	55.3	53.2
	param.	1.0M	2.1M	1.5M	2.1M	1.1M	2.1M	1.4M	2.1M	1.2M	1.9M	0.9M	1.9M
BP_2	WER	25.4	24.3	26.4	26.2	53.3	52.8	52.2	51.4	65.9	64.4	53.9	53.4
	param.	1.7M	2.6M	2.0M	2.5M	1.6M	2.6M	1.8M	2.6M	1.5M	2.4M	1.2M	2.4M
BP_3	WER	27.0	25.7	28.1	27.6	54.0	52.9	52.7	52.8	68.7	67.1	56.7	56.2
	param.	1.0M	1.6M	1.3M	1.7M	0.9M	1.6M	1.1M	1.6M	0.7M	1.2M	0.6M	1.2M

Table 3. Effect of type of training on WER (English-Spanish).

corpus alignment	estimation	normalization	equation	WER
sentence	EM	direct model	(6)	27.2
word	frequencies	standard approach	(7)	26.8
word	frequencies	direct model	-	24.3
word	EM	direct model	-	24.7

Table 4. Effect of type of search on WER.

	English Spanish	Spanish English	English French	French English	English German	German English
Best alignment	24.3	26.2	52.8	51.4	64.4	53.4
Add all alignments	24.8	26.5	52.6	52.0	64.2	53.4
Non-monotone search	24.1	26.2	53.1	51.3	64.4	53.1

Table 5. WER obtained in XRCE task by the models: GIATI [10] and PB.

	English Spanish	Spanish English	English French	French English	English German	German English
GIATI	28.7	31.8	61.7	56.6	66.9	61.3
PB	24.3	26.2	52.2	50.8	63.8	53.2

similarities: They are trained from word aligned corpus (obtained with GIZA++) and, at the first step, they extract a set of bilingual phrases from this word aligned corpus.

6 Conclusions

We have described a pattern recognition approach for performing statistical machine translation based on phrases. This approach is very simple and the search can be performed in reduced time (especially the monotone search). This method can obtain good translation results for certain tasks such as some reduced-domain tasks and can be extended easily to *computer-assisted translation*[13].

We have developed several experiments for the XRCE corpus of the TransType2 project. The main conclusions of our experiments are the following: Using long phrases reduces the WER. The best training is a direct approach using relative frequencies. Monotone and non-monotone searches obtain similar results.

The main contributions of this work are: i) A new formulation to PB model is proposed; ii) Several procedures for obtaining bilingual phrases are compared.

Acknowledgements

This work has been partially supported by the Spanish project TIC2003-08681-C02-02, by the IST Programme of the European Union IST-2001-32091 and by *Agencia Valenciana de Ciencia y Tecnología* under contract GRUPOS03/031.

References

1. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics **19** (1993) 263–311
2. Vogel, S., Ney, H., Tillmann, C.: HMM-based word alignment in statistical translation. In: COLING '96: The 16th Int. Conf. on Computational Linguistics, Copenhagen, Denmark (1996) 836–841
3. Tomás, J., Casacuberta, F.: Binary feature classification for word disambigaution in statistical machine translation. In: Procs. of the 2nd International Workshop on Pattern Recognition and Information Systems, Alicante, Spain (2002)
4. Och, F.J.: Statistical Machine Translation: From Single-Word Models to Alignment Templates. PhD thesis, Computer Science Dep., RWTH Aachen, Germany (2002)

5. Tomás, J., Casacuberta, F.: Monotone statistical translation using word groups. In: Procs. of the Machine Translation Summit VIII, Santiago, Spain (2001)
6. Zens, R., Och, F.J., Ney, H.: Phrase-based statistical machine translation. Advances in Artificial Inteligence **LNAI 2479** (2002) 18–32
7. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL), Edmonton, Canada (2003)
8. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29** (2003) 19–51
9. Tomás, J., Casacuberta, F.: Combining phrase-based and template-based models in statistical machine translation. In: Pattern Recogn. and Image Analisys. Volume 2652 of Lecture Notes in Computer Science. Springer-Verlag (2003) 1021–1031
10. Casacuberta, F.: Inference of finite-state transducers by using regular grammars and morphisms. In: Grammatical Inference: Algorithms and Applications. Volume 1891 of Lecture Notes in Computer Science. Springer-Verlag (2000) 1–14
11. Berger, A.L., Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Gillett, J.R., Kehler, A.S., Mercer, R.L.: Language translation apparatus and method of using context-based translation models. United States Patent, No. 5510981 (1996)
12. Tomás, J., Casacuberta, F.: Statistical machine translation decoding using target word reordering. In: Structural, Syntactic, and Statistical Pattern Recong. Volume 3138 of Lecture Notes in Computer Science. Springer-Verlag (2004) 734–743
13. TT2: Transtype2-computer-assisted translation (tt2). Technical report (2002) Information Society Technologies (IST) Programme. IST-2001-32091.
14. Zens, R., Ney, H.: Improvements in phrase-based statistical machine translation. In: Proceedings of the Human Language Technology Conference (HLT-NAACL), Boston, MA, USA (2004) 257–264

Automatic Segmentation of Bilingual Corpora: A Comparison of Different Techniques[*]

Ismael García Varea[1], Daniel Ortiz[2], Francisco Nevado[2],
Pedro A. Gómez[1], and Francisco Casacuberta[2]

[1] Dpto. de Informática
Universidad de Castilla-La Mancha, 02071 Albacete, Spain
ivarea@info-ab.uclm.es
[2] Dpto. de Sistemas Informáticos y Computación
Instituto Tecnológico de Informática
Univ. Politécnica de Valencia, 46071 Valencia, Spain

Abstract. Segmentation of bilingual text corpora is a very important issue to deal with in machine translation. In this paper we present a new method to perform bilingual segmentation of a parallel corpus, *SPBalign*, which is based on phrase-based statistical translation models. The new technique proposed here is compared with other two existing techniques, which are also based on statistical translation methods: the *RECalign* technique, which is based on the concept of recursive alignment, and the *GIATIalign* technique, which is based on simple word alignments. Experimental results are obtained for the EuTrans-I English-Spanish task, in order to create new, shorter bilingual segments to be included in a translation memory database. The evaluation of these three methods has been performed comparing the bilingual segmentations obtained by these techniques with respect to a manually segmented bilingual test corpus. These results show us that the new method proposed here outperforms in all cases the two already proposed bilingual segmentation techniques.

1 Introduction

During the last decade, the pattern recognition approach to machine translation, also known as statistical machine translation (SMT) [1], has been widely and successfully applied, obtaining better translation results than other more linguistically motivated approaches. However, SMT systems are far from being perfect, but can also be used in computer-assisted translation (CAT) in order to increase the productivity of the (human) translation process. Most of these systems are based on comparisons between a source sentence and reference sentences stored in translation memories (TMs). All of the systems based on TMs have a common principle: a text has many sentences which are similar to sentences that occur in

[*] This work has been partially supported by the Spanish CICYT under grant TIC2003-08681-C02-02, the *Agencia Valenciana de Ciencia y Tecnología* under contract GRUPOS03/031, the *Generalitat Valenciana*, and the project AMETRA (INTEK-CN03AD02).

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 614–621, 2005.

other texts and can be reused in new translations. Commercial systems usually do not use translation units that are shorter than a sentence. The translation search is done by similarity: the system is able to look for sentences in a database which are similar to the source sentence.

Typically, the degree of similarity of segments which are smaller than a sentence is higher than the similarity of complete sentences. Recent research works [2] have tried to work with TMs at a level which is smaller than the sentence level. A human translator usually decomposes the sentence to be translated and works with smaller units; therefore, it would be desirable to enrich the TM database with smaller translation units as well. Therefore, the automatic extraction of these new translation units in an efficient way, seems to be mandatory to develop competing TM systems. This is the main purpose of this work.

In this paper we present a new technique to perform bilingual segmentations of a parallel corpus, that we have called *SPBalign*. This method is a statistical phrase-based bilingual segmentation technique which is inspired in recently proposed phrase-based statistical translation models within the framework of SMT [3]. Our proposal is the application of three automatic *bilingual segmentation* (*bisegmentation*) techniques based on statistical translation methods to create new, shorter *bilingual segments* (*bisegments* or simply *biphrases*[1]).

The rest of the paper is organised as follows: In section 2 we briefly describe the already proposed *RECalign* and *GIATIalign* bisegmentation techniques, also the new bisegmentation technique proposed here is presented in a more detailed fashion. In section 3, the experimental bisegmentation results are presented for the EuTrans-I English-Spanish task [4]. In section 4, the conclusions of this work and future research directions are outlined.

2 Bilingual Segmentation

The purpose of a bisegmentation technique is to obtain translation units at a subsentence level. We redefine the formal definition of the bisegmentation concept presented in [5] as follows:

Let $f_1^J = f_1, f_2, \ldots, f_J$ be a source sentence and $e_1^I = e_1, e_2, \ldots, e_I$ the corresponding target sentence in a bilingual corpus. A bisegmentation S of f_1^J and e_1^I is defined as a set of ordered pairs included in $\mathcal{P}(f_1^J) \times \mathcal{P}(e_1^I)$, where $\mathcal{P}(f_1^J)$ and $\mathcal{P}(e_1^I)$ are the set of all subsets of consecutive sequence of words, of f_1^J and e_1^I, respectively. Each of the ordered pairs of the segmentation define a bisegment. In the following subsections we describe the three bisegmentation techniques that we have used in this work.

2.1 Recursive Bilingual Segmentation

Basically, a recursive alignment is an alignment between phrases of a source sentence and phrases of a target sentence. A recursive alignment represents the

[1] Here, the term *phrase* refers to a consecutive sequence of words, not necessarily with a linguistic structure or an independent meaning. In the following we will use the terms bisegment or biphrase indistinctly.

translation relations between two sentences, but it also includes information about the possible reorderings needed in order to generate the target sentence from the source sentence. A recursive alignment can be represented using a binary tree, where the internal nodes store the reordering directions and the leaf nodes store the translation relations. From a recursive alignment, a bisegmentation can be obtained by considering only the bisegments in the leaf nodes.

A greedy algorithm [6, 7] is used to compute recursive alignments from a bilingual corpus aligned at the sentence level. This greedy algorithm [6, 7] computes a recursive alignment for a source and a target sentence in this way: given the two sentences, it computes the most probable breakpoint in each sentence. Then, if the translation probability for the whole sentences is higher than the translation probability of dividing them, it creates a leaf node where the output sequence is considered to be the translation of the input sequence and it stops. Otherwise, it creates a new inner node of the tree and recursively applies the algorithm to the left and the right children. The bisegmentations are obtained as a byproduct from the recursive alignments.

This algorithm also uses the information of a statistical word alignment [8] between a pair of sentences in order to control the selection of the breakpoints in each sentence, that is, the breakpoints must be selected according to the relations that the word alignment imposes.

This system will be referred as *RECalign* in section 3.

2.2 GIATI-Based Bilingual Segmentation

The GIATI technique is an automatic method to infer statistical finite-state transducers as described in [9]. As a first step, this technique carries out a labelling of the words of the source sentence with the words of the output sentence from a word alignment between both sentences.

This kind of labelling can produce a bisegmentation if we consider that the bisegments are composed of the source words and their corresponding labels of target words. Basically, the method labels every source word with its connected target words except when a reordering is done in the alignment. In this case, the method groups all the necessary source and target words in order to consider the reordering inside the bisegment. This system will be referred as *GIATIalign* in section 3.

2.3 Statistical Phrase-Based Bilingual Segmentation

Statistical phrase-based translation models have been recently developed and used in order to improve existing statistical machine translation systems. These models are mainly based on bilingual phrase stochastic dictionaries which are automatically trained from a parallel corpus. Most of the used phrase-based translation models differ essentially in the way they construct such a bilingual phrase dictionary. The first method to obtain a statistical bilingual phrase dictionary, was proposed in [3], which is based on a corpus aligned at word level, where the corpus is aligned at word-level in both directions (source to target, and vice-versa) in order to improve results . The same approach has been used for other authors as [10–14]. In [15], in addition, tries are used in order to find a unique segmentation of the sentence pairs. On the other hand, in [16], the bilingual phrase dictionary is based on alignments at the phrase level, where the phrase correspondence is directly extracted from a bilingual corpus, by using a phrase-based joint probability model that directly generates both the source and target sentences.

In this approach, that we have called *SPBalign*, we use the phrase extraction method described in [3] by using a word-level aligned parallel corpus. The statistical phrase-based dictionary was constructed by training a translation model similar to the one proposed in [11], which can be estimated by relative frequency given the collected bilingual phrases $(\tilde{f}, \tilde{e})$ with the extraction method, that is:

$$p(\tilde{f}|\tilde{e}) = \frac{count(\tilde{f}, \tilde{e})}{\sum_{\tilde{f}'} count(\tilde{f}', \tilde{e})}$$

where, $count(\tilde{f}, \tilde{e})$ refers to the number of times the biphrase has been extracted from the training corpus.

Then, given a pair of sentences (f_1^J, e_1^I) and a word alignment between them, the *SPBalign* algorithm obtains the best segmentation in K bisegments $(1 \leq K \leq \min(J, I))$, and implicitly the best phrase-alignment $\tilde{a}_1^K$ between them. The phrase-alignment is defined as a one-to-one mapping between source and target segments. In essence, the *SPBalign* algorithm computes the phrase-based Viterbi alignment for a given sentence pair $\tilde{V}(\tilde{f}_j, \tilde{e}_i)$, obtaining a bisegmentation of the parallel corpus.

Basically, the algorithm works as follows:

1. For every possible $K \in \{1 \cdots \min(J, I)\}$
   (a) Extract all possible bisegmentations of size K according to the restrictions of $A(f_1^J, e_1^I)$.
   (b) Compute the probability of all possible phrase-level alignments $\tilde{a}_1^K = \tilde{a}_1 \cdots \tilde{a}_K$ of the current segmentation.
2. Return the segmentation $(\tilde{f}_j, \tilde{e}_i), j = \tilde{a}_i$ of highest probability.

More formally, the phrase-based Viterbi alignment of length K can be defined as follows:

$$\tilde{V}_K(\tilde{f}_{\tilde{a}_i}, \tilde{e}_i) = \arg\max_{\tilde{a}_1^K} \left\{ \prod_{i=1}^{K} p(\tilde{f}_{\tilde{a}_i}|\tilde{e}_i) \right\}$$

The results obtained with this new bisegmentation method are shown in section 3.

3 Experimental Results

The bisegmentations obtained with the three studied techniques are compared with a reference bisegmentation computed manually by experts. In order to evaluate the bisegmentations achieved by these techniques, we will use the method described in [5, 7, 17]. Results are presented using the Recall (percentage of covered segmentations), Precision (percentage of perfect segmentations) and F-measure [5]. The F-measure is the *harmonic mean* of precision and recall, then, this measure give us a compromise between the coverage and exactness of the automatic obtained bilingual segmentation. These measures are defined as:

$$\text{Recall} = \frac{|S \cap S_r|}{S_r}; \quad \text{Precision} = \frac{|S \cap S_r|}{S}; \quad \text{F} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

where S is an obtained bisegmentation and S_r is a reference bisegmentation for a given bilingual corpus.

In order to present a detailed experimentation, for each technique presented above, we carried out different experiments according to the type of word alignment that was used to bisegment the test corpus. That is, the source-to-target word alignment, and also the target-to-source word alignment, (E-S) for English-to-Spanish and (S-E) for Spanish-to-English. Moreover, three different combinations of both alignments were used: the intersection ($\cap$), the union ($\cup$) , and the refined (R) symmetrization methods proposed in [3]. All the word alignments used in the experimentation process were carried using the GIZA++ toolkit [18].

3.1 Corpus Description

For the experiments, we have used the EUTRANS-I corpus, which is a Spanish-English bilingual corpus whose domain is a subset of the tourist task [4]. From this corpus, 10,000 different sentence pairs were selected for training purposes. We also selected a test corpus, not included in the training corpus, consisting on 40 randomly selected pair of sentences. The 40-sentences test corpus was bilingually segmented by human experts. Table 1 shows the characteristics of the training and test sets for this corpus.

Table 1. Training and Test sets of the EUTRANS-I corpus.

	Training		**Test**	
	Spanish	English	Spanish	English
Sentences	10,000		40	
Words	97,131	99,292	491	487
Vocabulary	686	513	149	126
Trigram Perplexity	–	–	4.6	3.6

3.2 Bisegmentation Quality Results

The bisegmentation results for the Spanish-English and the English-Spanish translation directions are presented in tables 2 and 3, respectively. The four different types of word alignment are used with every technique. For every technique, the best result are highlighted in bold.

With independence of the technique used, the more restrictive alignment information the better results are obtained. This is performed using the union word

Table 2. Bisegmentation results for Spanish-English translation direction.

Technique	**Recall**	**Precision**	**F-measure**
RECalign+(S-E)	39.67	87.11	54.51
RECalign+($\cap$)	36.60	87.42	51.60
RECalign+($\cup$)	52.96	79.01	**63.41**
RECalign+(R)	48.86	80.67	60.85
GIATIalign+(S-E)	39.91	85.92	54.50
GIATIalign+($\cap$)	36.22	80.26	49.91
GIATIalign+($\cup$)	39.99	85.52	**54.50**
GIATIalign+(R)	37.35	84.68	51.84
SPBalign+(S-E)-5	68.09	68.47	68.28
SPBalign+($\cap$)	67.21	67.38	67.29
SPBalign+($\cup$)	72.58	65.49	**68.85**
SPBalign+(R)	66.27	65.84	66.06

Table 3. Bisegmentation results for English-Spanish translation direction.

Technique	**Recall**	**Precision**	**F-measure**
RECalign+(E-S)	50.68	71.75	59.40
RECalign+($\cap$)	50.11	71.61	58.96
RECalign+($\cup$)	62.92	66.11	**64.47**
RECalign+(R)	57.71	68.84	62.79
GIATIalign+(E-S)	40.64	81.19	54.16
GIATIalign+($\cap$)	36.48	73.96	48.86
GIATIalign+($\cup$)	41.21	82.69	**55.01**
GIATIalign+(R)	40.40	82.34	54.21
SPBalign+(E-S)	73.39	61.08	66.67
SPBalign+($\cap$)	72.63	57.57	64.23
SPBalign+($\cup$)	67.75	65.88	**66.80**
SPBalign+(R)	67.75	65.88	66.80

alignment for both translation directions. That is exactly what we expected, because the alignment union remarks the alignments between words which are good translations of each other, which finally results in a better bisegmentation, and, implicitly the alignment at segment level.

In general, similar accuracy is obtained in both translation directions. In all cases the *SPBalign* technique proposed here, outperforms the *RECalign* and *GIATIalign* techniques.

Regarding the results with respect to precision and recall, we can see that the *SPBalign* technique obtains a more balanced values of both measures. In most of the cases, the F-measure results of the other two techniques are biased by the precision values, that is, these techniques are very precise but paying the cost of not performing a good coverage of bilingual segments. In contrast, the *SPBalign* is not so precise as the others but the results of this technique are not biased by the precision values nor by the recall ones.

4 Conclusions and Future Work

A new automatic bilingual segmentation technique, namely *SPBalign*, inspired in statistical phrase-based translation models has been presented. This technique has been compared to other two previously proposed techniques: one based on the concept of recursive alignment (*RECalign*) and another one based on simple word alignments (*GIATIalign*).

According to the experimental results presented here, the *SPBalign* technique, outperforms the other two, and it also achieves a better coverage than the others. Regarding the comparison of the three techniques we can conclude that:

- In general, for the three studied techniques, better results are obtained when more word-alignment information is used.
- Similar results are obtained in both translation directions. We think that it is due to the similarity between the source and target languages, so we cannot conclude that this fact will be true for more distant language pairs, as for example Spanish and Basque.

For the future we have planned:

- To explore new symmetrization methods to combine alignments at word level in order to obtain better results, as for example the use of the union of a list of n-best word alignments.
- In the same direction, we think that it could be also useful to use a sort of weighted word-alignments in order to focus only on those word alignments that we are sure to be correct.
- We are thinking to extend the *SPBalign* technique taking into account the information of the probability of a specific segmentation, which can be also obtained from the training of the phrase-based translation models.
- To apply the different bisegmentation techniques to more complex translation tasks, as HANSARDS or VERBMOBIL tasks.
- To evaluate these techniques by means of translation quality, instead of alignment quality. This will give us a more realistic information about the usability of such bisegmentation methods.

References

1. Ney, H.: Stochastic modelling: From pattern classification to speech recognition and translation. In: Proceedings of the International Conference on Pattern Recognition, Barcelona, Spain, IAPR (2000) 3025–.3032
2. Simard, M., Langlais, P.: Sub-sentential exploitation of translation memories. In: Proc. of MT Summit VIII. (2001)
3. Och, F.J.: Statistical Machine Translation: From Single-Word Models to Alignment Templates. PhD thesis, Computer Science Department, RWTH Aachen, Germany (2002)

4. Amengual, J., Benedí, J., Casacuberta, F., no, M.C., Castellanos, A., Jiménez, V., Llorens, D., Marzal, A., Pastor, M., Prat, F., Vidal, E., Vilar, J.: The EuTrans-I speech translation system. Machine Translation **1** (2000)
5. Simard, M., Plamondon, P.: Bilingual sentence alignment: Balancing robustness and accuracy. Machine Translation **13** (1998) 59–80
6. Nevado, F., Casacuberta, F.: Bilingual corpora segmentation using bilingual recursive alignments. In: Proceedings of the 3th Jornadas en Tecnologías del Habla, Valencia (2004)
7. Nevado, F., Casacuberta, F., Landa, J.: Translation memories enrichment by statistical bilingual segmentation. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon (2004)
8. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics **19** (1993) 263–311
9. Casacuberta, F., Vidal, E.: Machine translation with inferred stochastic finite-state transducers. Computational Linguistics **30** (2004) 205–225
10. Zens, R., Och, F., Ney, H.: Phrase-based statistical machine translation. In: Advances in artificial intelligence. 25. Annual German Conference on AI. Volume 2479 of Lecture Notes in Computer Science. Springer Verlag (2002) 18–32
11. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL), Edmonton, Canada (2003)
12. Tillmann, C.: A projection extension algorithm for statistical machine translation. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP2003). (2003) 1–8
13. Venugopal, A., Vogel, S., Waibel, A.: Effective phrase translation extraction from alignment models. In: Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL). (2003) 319–326
14. Vogel, S., Zhang, Y., Huang, F., Tribble, A., Venugopal, A., Zhao, B., Waibel, A.: The CMU statistical machine translation system. In: Proc. of Machine Translation Summit IX, New Orleans, USA (2003) 115–120
15. Zhang, Y., Vogel, S., Waibel, A.: Integrated phrase segmentation and alignment algorithm for statistical machine translation. In: International conference on natural language processing and knowledge engineering, Beijing, China (2003)
16. Marcu, D., Wong, W.: A phrase-based, joint probability model for statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002). (2002)
17. Langlais, P., Simard, M., Véronis, J.: Methods and practical issues in evaluating alignment techniques. In: Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th Int. Conf. on Computational Linguistics, Montreal (1998) 717–717
18. Och, F.J., Ney, H.: Improved statistical alignment models. In: Proc. of the 38th Annual Meeting of the Association for Computational Linguistics (ACL), Hong Kong, China (2000) 440–447

Word Translation Disambiguation Using Multinomial Classifiers[*]

Jesús Andrés, José R. Navarro, Alfons Juan, and Francisco Casacuberta

Departamento de Sistemas Informáticos y Computación
Instituto Tecnológico de Informática
Universidad Politecnica de Valencia

Abstract. This work focuses on a hybrid machine translation system from Spanish into Catalan called SisHiTra. In particular, we focus on its word translation disambiguation module, which has to decide on the correct translation of each ambiguous input word in accordance with its context. We propose the use of statistical pattern recognition techniques for this task and, in particular, multinomial Naive Bayes text classifiers. Extensive empirical results on the use of these classifiers are presented, in which the influence of the window (context) size and parameter smoothing are carefully studied.

1 Introduction

Machine translation is a hot research area due to its increasingly important economic and social impact. Knowledge-based, corpus-based and hybrid systems are being successfully developed for many pairs of languages, especially in the case of domain-specific tasks. Open-domain translation, however, is still a difficult challenge for most pairs of languages, though very good results are being obtained for pairs of closely-related languages. An example of this is Spanish and Catalan, two romaninc languages which have many characteristics in common.

This work focuses on a hybrid translation system from Spanish into Catalan called *SisHiTra* (from the Spanish "Sistemas Híbridos de Traducción", Hybrid Translation Systems) [2]. SisHiTra combines linguistic knowledge and statistical pattern recognition techniques in order to provide accurate, open-domain translation from Spanish into Catalan. The performance of SisHiTra is similar or better than other translation systems such as SALT [8], and Internostum [9] (see [2]). However there is still room for improvement in certain modules of the system and, in particular, in its *word translation disambiguation* module. This module uses a bilingual dictionary with all the possible translations of different meaning each Spanish word has; so, in accordance with this dictionary, words with two or more possible translations are ambiguous, and hence they have to be disambiguated whenever they appear in an input text. It is assumed that the context of an ambiguous word carries enough information to perform

[*] Work supported by the "Agència Valenciana de Ciència i Tecnologia" under grant GRUPOS03/031 and the Spanish project ITEFTE (TIC2003-08681-C02-02).

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 622–629, 2005.

disambiguation reliably, and this is precisely the goal of the word translation disambiguation module. Unfortunately, it is not straightforward to derive accurate knowledge-based rules for all words in all possible contexts. In fact, the current version of the module ignores context and has a predefined answer for each word.

This paper explores the idea of using statistical pattern recognition techniques for word translation disambiguation and, in particular, *multinomial Naive Bayes text classifiers* [1, 6]. Our work is, to some extent, a continuation of the work reported in [5], where limited yet promising results are obtained using similar classifiers. Here we present extensive empirical results on the use of multinomial classifiers for word translation disambiguation, in which the influence of the *window (context) size* and *parameter smoothing* are carefully studied. These results are given in section 4, after a brief revision of SisHiTra and multinomial classifiers in the two following sections.

2 SisHiTra

Given an input sentence in Spanish, SisHiTra translates it into Catalan by executing seven basic steps (modules). In broad terms, these steps can described as follows (see [2] for a detailed description):

1. *Fragmentation:* the input text (in Spanish) is divided into *sentences*
2. *POS-tagging:* a Part-Of-Speech tag is chosen for each input word [4]
3. *Nominal phrase agreement:* nominal phrases are located in the input and corrected for gender and number agreement
4. *Localisation:* each input word or expression is labelled with all of its possible translations
5. *Word disambiguation in translation:* a unique translation is chosen for each input word or expression
6. *Inflection and Formatting:* translations are inflected in Catalan and contraction and apostrophization rules for Catalan are applied
7. *Integration:* translated fragments are compiled in accordance with the original input format

In this paper, the problem we are interested in is word translation disambiguation. Consider, for instance, the Spanish word *en*, which has four different translations in Catalan: *en*, *a*, *per* and *amb*. In *"Era un ejecutivo* ***en*** *viaje de negocios que regresaba a casa"*, its correct translation is also *en: "Era un executiu* ***en*** *viatge de negocis que tornava a casa"*. However, in *"Tenía el honor de ser el primer vuelo retrasado del año* ***en*** *Valencia"*, it means *a* in Catalan: *"Tenia l'honor de ser el primer vol endarrerit de l'any* ***a*** *València"*. It is not difficult to find examples in which *en* translates into *per* or *amb* in Catalan and, in fact, neither of the four possible translations is extremely rare. Therefore, it is not a good idea to always give the same answer, as SisHiTra currently does. Instead, if we assume that the correct translation depends on the *context* (surrounding words), then it is better to learn such dependency and use it to minimise the probability of disambiguation error.

3 Multinomial Classifiers for Disambiguation

Word translation disambiguation of each (ambiguous) word in the input (Spanish) vocabulary entails a separate classification problem. Let D be the input vocabulary size and let C_d be the number of different translations for word d, $d = 1, \ldots, D$. (Note that, for simplicity, we are also including the non-ambiguous case of $C_d = 1$.) In probabilistic terms, our basic problem is to find the most probable translation for each word d, in each prefix-suffix context (u, v), that is,

$$c^*(d; u, v) = \underset{c=1,\ldots,C_d}{\arg\max}\ p(c \mid d, u, v) \tag{1}$$

Clearly, there are too many possible contexts to consider, and hence we will not be able to accurately estimate the *posterior* probabilities in (1) for each possible context. To make things simpler, let us assume that these probabilities do not depend on the *order* in which words appear in the context, but only on which words occur and how many times. This leads to the well-known "bag of words" representation for text; i.e., in our case, every context (u, v) is represented as a D-dimensional vector of word counts, $\mathbf{x} = (x_1, \ldots, x_D)^t$, with x_d being the number of occurrences of word d in u and v. Also, in combination with this assumption, we may limit the context of a word to its immediately surrounding neighbours, thus assuming that words farther away in the sentence do not have significant influence on its disambiguation. With these assumptions in mind, our basic problem can be rewritten as:

$$\begin{aligned} c^*(d; u, v) &\approx \underset{c=1,\ldots,C_d}{\arg\max}\ p(c \mid d, \mathbf{x}) && (2)\\ &= \underset{c=1,\ldots,C_d}{\arg\max}\ \log p(c \mid d) + \log p(\mathbf{x} \mid d, c) && (3) \end{aligned}$$

by the Bayes' rule and the increasing monotonicity of log.

Now we make the Naive Bayes assumption that words in a context $\mathbf{x}$ occur with independence of each other though, of course, their probabilities of occurrence depend on the word d and its translation c that are being considered. Then, the word-and-class conditional probability in (3) is a multinomial p.f.,

$$p(\mathbf{x} \mid d, c) = \frac{x_+!}{\prod_{d'} x_{d'}!} \prod_{d'} p(d' \mid d, c)^{x_{d'}} \tag{4}$$

where x_+ is the total number of words in context $\mathbf{x}$ ($x_+ = \sum_{d'} x_{d'}$) and $p(d' \mid d, c)$ is the (unknown) probability of word d' to occur in a context of word d with translation c. (For (4) to define a proper probability distribution over all possible contexts, we must only consider all contexts having equal x_+; otherwise, it is still applicable but improper.) Plugging (4) into (3), we get

$$\begin{aligned} c^*(d; u, v) &\approx \underset{c=1,\ldots,C_d}{\arg\max}\ \log p(c \mid d) + \log \frac{x_+!}{\prod_{d'} x_{d'}!} + \sum_{d'} x_{d'} \log p(d' \mid d, c) && (5)\\ &= \underset{c=1,\ldots,C_d}{\arg\max}\ \log p(c \mid d) + \sum_{d'} x_{d'} \log p(d' \mid d, c) && (6) \end{aligned}$$

which, for each word d, is a log-linear classifier depending on the vector of unknown parameters (probabilities) $\mathbf{\Theta}_d = \{p(c \mid d), p(d' \mid d, c)\}_{d',c}$.

Maximum likelihood estimation of $\mathbf{\Theta}_d$ w.r.t. a random sample from $p(\mathbf{x}, c \mid d)$ leads to the conventional estimates for *priors* and multinomial parameters:

$$\hat{p}(c \mid d) = \frac{N(d,c)}{\sum_{c'} N(d,c')} \tag{7}$$

$$\hat{p}(d' \mid d, c) = \frac{N(d',d,c)}{\sum_{d''} N(d'',d,c)} \tag{8}$$

where $N(d,c)$ is the number of training contexts of word d with translation c, and $N(d',d,c)$ is the number of occurrences of word d' in contexts of word d with translation c.

Unfortunately, the probability estimates given by (8) often underestimate the true probabilities involving rare word-translation pairs. For instance, in one on the experiments reported in section 4, we face the problem of computing 320 millions of such probability estimates and only 1.3 millions of them are non-zero. To circumvent this problem, we have considered two basic parameter smoothing techniques [1, 6] which are known to give good results in similar text classification problems. The first of these techniques, known as *Laplace smoothing*[1], consists in adding a "pseudo-count" $\epsilon > 0$ to every $N(d',d,c)$ count:

$$\tilde{p}(d' \mid d, c) = \frac{N(d',d,c) + \epsilon}{\sum_{d''} N(d'',d,c) + \epsilon} \tag{9}$$

with $\epsilon = 1$ as default value.

The second smoothing technique we have considered is *absolute discounting* [1, 6]. Instead of using artificial pseudo-counts, we gain "free" probability mass by discounting a small constant b, $0 < b < 1$, to every word d' with non-null count $N(d',d,c)$. The gained probability mass is then distributed among words with null counts (*backing-off*), or all words (*interpolation*), in accordance with a *generalised distribution* such as a *uniform* or *unigram* distribution [1].

Parameter smoothing is needed to prevent null probability estimates, especially in the case of contexts of limited, normalised length (e.g. one word at both sides of the word being disambiguated). In the opposite case, i.e. in the case of contexts with no length limits, every word in a sentence counts in the context of every other word in the same sentence, and hence there are less chances of getting null probability estimates according to (8). In this case, however, we may have a different problem of bias towards correctly disambiguating words in long contexts at the expense of short ones. This is due to the fact that we are not modelling context length, and thus the estimates given by (8) are clearly dominated by counts coming from long contexts. As discussed in [1], a practical remedy to this problem is to normalise the word counts of each context with respect to its length:

$$\tilde{\mathbf{x}} = L \, \frac{\mathbf{x}}{x_+} \tag{10}$$

where L is any convenient normalisation constant such as the average sentence length or simply 1.

4 Experiments

Experiments are based on a parallel Spanish-Catalan corpus extracted from the newspaper *El Periódico* [7]. This corpus consists of $806K$ parallel sentences comprising of 28.6M running words (13.7M Spanish and 14.9M Catalan). It was automatically aligned at word level using the *GIZA++* software tool [3].

We have used a dictionary with 7,085 ambiguous words included in SisHiTra. From them, 2,207 do not occur in the corpus, 460 occur only once, and the remaining 4,418 occur twice or more. For each ambiguous word occurring twice or more, the error rate of the multinomial classifier was estimated using a 2-fold cross validation procedure on all of its contexts found in the corpus. This was done for several smoothing techniques (Laplace and four variants of absolute discounting) and for context window sizes of 1, 2, 3 and ∞ (number of words at both sides of the word being disambiguated that are included in its context). Also, in the case of no window size limit (∞), it was done for both unnormalised and normalised word counts. Figure 1 shows the average error (percentage of misclassified contexts) obtained in these tests.

The results in Figure 1 suggest that, in general, a local window size (of 2, for instance) and Laplace smoothing (or absolute discounting with uniform generalised distribution) are appropriate for word translation disambiguation in our task. The smoothing discount does not have an important effect since the curves are almost flat in all cases. Also note that context length normalisation does not have a significant influence in the case of contexts with no size limits.

Clearly, there is no need to use the very same window size and smoothing technique (and discount) for all ambiguous words in an optimised multinomial-based disambiguator for SisHiTra. Then, it is more realistic to provide results using the best window size, smoothing technique (and discount), or both, for each word in training. This is done in Tables 1 and 2, where the multinomial disambiguator is compared with the rule-based technique currently used in SisHiTra, and also with a maximum a priori probability classifier (i.e., a classifier that chooses the class having maximum a priori probability, as given in Eq. 7).

Table 1 reports the best results obtained as a function of window size (so, adjusting the smoothing technique and discount for each word). The best result, 3.5%, corresponds to a window size of 2. It compares favourably with the 9.6% yielded by SisHiTra and also with the 6.6% obtained by maximum a priori probability classification. It can be slightly reduced to a 3.4% of error by letting each word have its best window size for disambiguation.

Table 2 includes specific results for the 10 most frequent Spanish words in the corpus. There are four (almost) non-ambiguous words for which we get no error: *de*, *a*, *tener* and *como*. Similarly, *que* and *ser* give small error. The most interesting cases correspond to the remaining four words: *en*, *para*, *le* and *lo*. In *en*, *le* and *lo*, the multinomial disambiguator gives much better results than SisHiTra and the maximum a priori probability classifier. In the case of *para*, SisHiTra gives a 94.7% of error while the other two improve this figure up to a 5.3%. We have further analysed this case and found that SisHiTra always translates *para* into *per a*, but its most probable translation in the corpus is *per*.

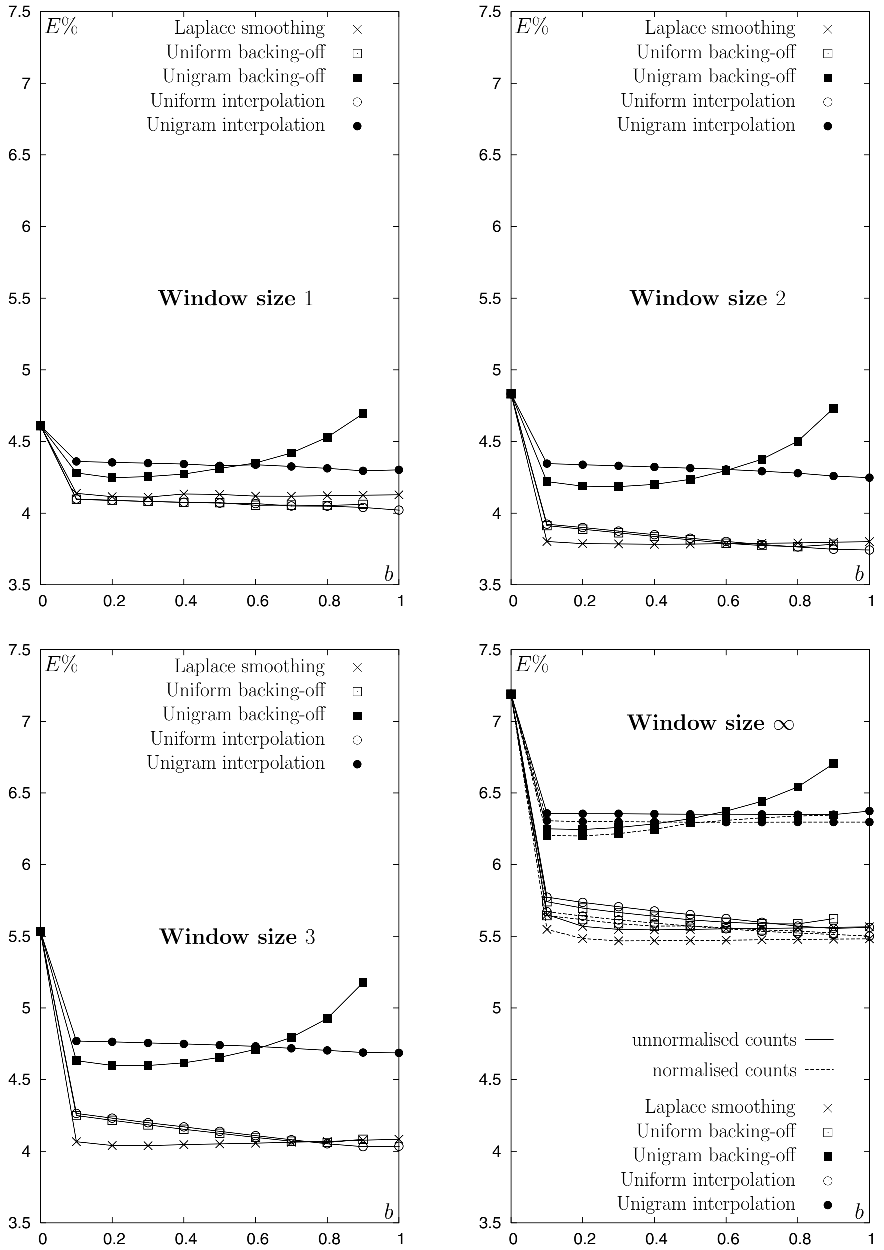


Fig. 1. Average error rate (percentage of misclassified contexts) of the multinomial classifier, as a function of the smoothing discount, for several smoothing techniques, window sizes (one in each panel) and, in the case of no window size limit (∞), for both unnormalised and normalised word counts.

Table 1. Best average error rate (percentage of misclassified contexts) for several window sizes: -1 = SisHiTra, 0 = maximum a priori probability classifier, $n[\geq 1]$ = multinomial classifier with window size of n.

Window size	Error (%)
-1	9.6
0	6.6
1	3.9
2	3.5
3	3.8
∞	5.4

Table 2. Error rates for some ambiguous words given by SisHiTra, and the maximum a priori and multinomial classifiers.

		SisHiTra	A priori	Multinomial			
Word	Freq.%	Error%	Error%	Error%	Smoothing	b	Win. size
de	28.9	0.0	0.0	0.0	-	-	3
en	9.6	42.2	42.2	17.4	-	-	2
a	7.7	0.0	0.0	0.0	-	-	3
que	5.4	1.6	1.6	1.4	Abs. Discount	0.1	2
ser	4.8	0.3	0.3	0.3	Abs. Discount	0.9	1
para	2.5	94.7	5.3	5.3	Laplace	1.0	∞
tener	1.0	0.0	0.0	0.0	-	-	3
le	0.8	12.9	12.9	9.7	-	-	1
como	0.7	0.0	0.0	0.0	-	-	3
lo	0.7	36.9	36.9	20.1	Abs. Discount	1.0	1

A new disambiguation module for SisHiTra was developed in accordance with the empirical results described above. More precisely, the module was built from an optimised multinomial classifier for each ambiguous word, trained from all contexts found in the corpus and the best design parameter values tried in the experiments. The new module was empirically compared with the previous knowledge-based module on two tests: the first is a hard test involving 511 difficult sentences, plenty of ambiguous words (about 3K occurrences of ambiguous words out of 9K running words); the second test is a random set from “El Periódico” of 3,560 sentences (about 13,500 occurrences of ambiguous words out of 52,700 running words). Results, both in terms of WER[1] (*Word Error Rate*) and CER[2] (*Classification Error Rate*) are given in Table 3. From these results, we may conclude that the new multinomial-based disambiguation module performs better than its predecessor.

[1] WER:The minimum number of substitution, insertion and deletion operations needed to convert the word string hypothesised by the translation system into a given single reference words.

[2] Of course, Classification Error Rate over ambiguous words.

Table 3. Word Error Rate (WER) and Classification Error Rate (CER) of ambiguous words for the knowledge-based and multinomial disambiguator, tested on a hard ambiguous test and a set of sentences from "El Periódico".

Disambiguator	Ambiguous Test		"El Periódico"	
	WER	CER	WER	CER
Knowledge-based	20.7	30.0	25.7	27.0
Multinomial	19.5	26.7	24.7	23.2

5 Conclusions

The multinomial Naive Bayes text classifier has been studied for word translation disambiguation in a hybrid machine translation system from Spanish to Catalan. Empirical results have been reported in which the influence of the window (context) size and parameter smoothing have been carefully studied. The new multinomial-based disambiguation module performs better than its knowledge-based predecessor.

The main direction of our current work is to compare this new module with other statistical approaches and, in particular, with classifiers based on conventional language models such smoothed trigrams. In fact, we have already done preliminary experiments in this direction, but the results obtained so far are not satisfactory.

References

1. Alfons Juan and Hermann Ney. Reversing and Smoothing the Multinomial Naive Bayes Text Classifier. In *Proc. of the 2nd Int. Workshop on Pattern Recognition in Information Systems (PRIS 2002)*,pages 200–212,Alicante (Spain), April 2002.
2. José R. Navarro and Others. SisHiTra: A Hybrid Machine Translation System from Spanish to Catalan. In *Proceedings of the ESTAL 04 Workshop on Advances in Natural Language Processing*,Lecture Notes in Artificial Intelligence,pages 349–359. Springer-Verlag, 2004.
3. F. J. Och and H. Ney. Improved Statistical Alignment Models. In *ACL00*, pages 440–447, Hongkong, China, October 2000.
4. E. Roche and Y. Schabes. Deterministic Part-Of-Speech Tagging with Finite-State Transducers. In *Computational Linguistics*,21(2):pages 227–253, 1995.
5. J. Tomás and F. Casacuberta. Binary deature classification for word disambiguation in statistical machine translation. In *Proceedings of the 2nd International Workshop on Pattern Recognition in Information Systems*, pages 213–224, Spain, 2002.
6. David Vilar and Hermann Ney and Alfons Juan and Enrique Vidal. Effect of Fature Smoothing Methods in Text Classification Tasks. in *Proc. of the 2nd Int. Workshop on Pattern Recognition in Information Systems (PRIS 2004)*, 2004.
7. "El Periódico" website: www.elperiodico.com. Ediciones Primera Plana S.A., Consell de Cent, 425-427. 08009 Barcelona (Spain)
8. *http://www.cult.gva.es/salt/salt_programes_salt2.htm*
9. *http://www.internostrum.com/*

Different Approaches to Bilingual Text Classification Based on Grammatical Inference Techniques[*]

Jorge Civera[1], Elsa Cubel[2], Alfons Juan[1], and Enrique Vidal[2]

[1] Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
{jcivera,ajuan}@dsic.upv.es

[2] Instituto Tecnológico de Informática, Universidad Politécnica de Valencia
{ecubel,evidal}@iti.upv.es

Abstract. Bilingual documentation has become a common phenomenon in many official institutions and private companies. In this scenario, the categorization of bilingual text is a useful tool, that can be also applied in the machine translation field. To tackle this classification task, different approaches will be proposed. On the one hand, two finite-state transducer algorithms from the grammatical inference domain will be discussed. On the other hand, the well-known naive Bayes approximation will be presented along with a possible modelization based on n-gram language models. Experiments carried out on a bilingual corpus have demonstrated the adequacy of these methods and the relevance of a second information source in text classification, as supported by classification error rates. Relative reduction of 29% with respect to the best previous results on the monolingual version of the same task has been obtained.

1 Introduction

Nowadays the proliferation of bilingual documentation is a widely extended phenomenon in our information society. This fact is reflected in a vast number of official institutions (EU parliament, the Canadian Parliament, UN sessions, Catalan and Basque Parliaments in Spain, etc.) and private companies (user's manuals, newspapers, books, etc.). In many cases, this textual information needs to be categorized by hand, entailing a time-consuming and arduous burden.

On the other hand, the categorization of bilingual text can be applied to the field of machine translation to train specific statistical translation models for particular subdomains, automatically detected in more general bilingual corpus. This strategy can alleviate a typical problem of statistical translation models: their limited application to constrained semantic scopes. In its monolingual form, text classification has demonstrated reasonably good performance using a number of well-known techniques such as naive Bayes [1], support vector machines [2],

[*] Work supported by the "Agència Valenciana de Ciència i Tecnologia" under grant GRUPOS03/031 and the Spanish project TIC2003-08681-C02-02.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 630–637, 2005.

k-nearest neighbours [3], etc. The incorporation of a second language offers an additional information source that can help reducing classification error rates.

In the present work, several techniques for classifying bilingual text will be presented along with some preliminary experimental results in a four-class bilingual limited-domain corpus. The rest of the paper is organized as follows. Next section will introduce the statistical classification model employed for bilingual classification. In Section 3, an instantiation of this classification model as two different stochastic finite-state transducer models will be considered. Alternatively, a naive Bayes approach to bilingual classification will be explained in Section 4. Finally, Section 5 will be devoted to experimental results and Section 6 will discuss some conclusions and future work.

2 Bilingual Classification

The bilingual classification task can be seen from a probabilistic perspective. Given a bilingual sample (s,t) composed by a pair of sentences (being t a possible translation for s) and a set of classes $1,2,\ldots,C$, the pair (s,t) will be assigned to class $\hat{c}$ according to the maximum *a posteriori* probability criterion:

$$\hat{c} = \underset{c}{\operatorname{argmax}}\, P(c \mid (s,t)) = \underset{c}{\operatorname{argmax}}\, p((s,t) \mid c)P(c) \tag{1}$$

This joint class-conditional probability can be modeled in various ways, some of them will be analyzed in the next sections.

3 Transducer Inference and Decoding Algorithm

Stochastic finite-state transducers (SFSTs) are translation models [4, 5] that can be learned automatically from bilingual sample pairs. SFSTs are finite-state networks that accept sentences from a given input language and produce sentences of an output language. Every edge of the network has associated an input symbol, an output string and a transition probability associated to the input-output pair. Every time an input symbol is accepted, the corresponding string is output and a new state is reached. Once the input sentence has been parsed completely, additional output may be produced from the last state reached.

Optimal maximum likelihood estimates of the transition probabilities can be obtained by computing the relative frequency each edge is used in the parsing of the input-output sample pairs. This results in SFSTs which models the joint probability distribution over bilingual sentence pairs derived in previous section. There exist several techniques that infer SFSTs efficiently. SFSTs have been successfully applied into several translation tasks [6, 7].

3.1 OSTIA

A particular case of SFST are known as subsequential transducers (SSTs). These are essentially SFSTs with the restriction of being deterministic. Given a set of training pairs of sentences from a translation task, the *Onward Subsequential*

Transducer Inference Algorithm (OSTIA) efficiently learns a SST that generalizes the training set [8]. The algorithm builds a straightforward prefix-tree representation of all the training pairs and moves the output strings towards the root of this tree as much as possible, leading to an "onward" tree representation. Finally a state merging process is carried out. The algorithm guarantees identification of total subsequential functions in the limit [9], that is, if the unknown target translation exhibits a subsequential structure, convergence to it is guaranteed whenever the set of training samples is representative.

Nevertheless, there are *partial* subsequential functions for which OSTIA inference is troublesome. This limitation can be solved by an extension, called OSTIA-DR (OSTIA with Domain and Range constraints) [10] in which the learnt transducers only accept input sentences and only produce output sentences compatible with the input/output language models.

Another possibility to overcome the partial-function limitation is to rely on statistical knowledge. Following this idea, an extension of OSTIA-DR is proposed, referred to as OSTIA-P [11].

OSTIA-P infers a *Stochastic Subsequential Transducer*, i.e., a SST with probabilistic information. The algorithm is based on the same state merging approach as in the basic OSTIA(-DR). In addition to the OSTIA(-DR) merging conditions, here two states are considered compatible for merging if the probabilities of their suffixes are similar within a certain threshold α. This threshold indirectly controls the level of generalization of the inferred stochastic finite-state models.

OSTIA-P is based on a more restrictive merging criterion than that of the original OSTIA(-DR). As a consequence, transducers generated by OSTIA-P will tend to reduce the generalization of the training data.

3.2 GIATI

Another possible algorithm for learning SFSTs is the *Grammatical Inference and Alignments for Transducer Inference* (GIATI) technique [4]. Given a finite sample of string pairs, it works in three steps:

1. Building training strings. Each training pair is transformed into a single string from an extended alphabet to obtain a new sample of strings.
2. Inferring a (stochastic) regular grammar. Typically, a smoothed n-gram is inferred from the set of samples of strings obtained in the previous step.
3. Transforming the inferred regular grammar into a transducer. The symbols associated to the grammar rules are replaced by source/target symbols, thereby converting the grammar inferred in the previous step into a transducer.

The transformation of a parallel corpus into a corpus of single sentences is performed with the help of statistical alignments: each word is joined with its translation in the output sentence, creating an "extended word". This joining is done taking care not to invert the order of the output words. The third step is trivial with this arrangement. In our experiments, the alignments are obtained using the GIZA software [12], which implements IBM statistical models [13].

3.3 Viterbi Error-Correcting Decoding

The performance achieved by transducer models tends to be poor if input samples do not strictly comply with the syntactic restrictions imposed by the model. This is the case of syntactically incorrect sentences, or correct sentences whose precise "structure" has not been exactly captured by the model during the training process.

Both of these problems can be approached by means of error-correcting decoding. Under this approach, the input sentence, x, is considered to be a *corrupted* version of some sentence, $\hat{x} \in \mathcal{L}$, where $\mathcal{L}$ is input language associated with the SFST. On the other hand, an error model $\mathcal{E}$ accounts for the transformation from $\hat{x}$ into x. In the current work, this error model is the edit cost model, that considers the transformation between two sentences $\hat{x}$ and x, in terms of edit word operations (insertions, deletions and substitutions). In practise, these errors should account for likely vocabulary variations, word dissapearances, superfluous words, repetitions, and so on.

Formalising the framework given above, our goal is to obtain a sentence $\hat{x}$ whose *a posteriori* probability of being generated from x is maximum:

$$\hat{x} = \operatorname*{argmax}_{x'} P(x'|x, \mathcal{L}, \mathcal{E}) \approx \operatorname*{argmax}_{x'} P(x|x', \mathcal{E}) \cdot P(x'|\mathcal{L}) \tag{2}$$

Under the classification framework, Eq. 2 is integrated in Eq. 1 as follows:

$$\hat{c} = \operatorname*{argmax}_{c} P(x|c) = \operatorname*{argmax}_{c} P(\hat{x}_c|c) \approx \operatorname*{argmax}_{c} P(\hat{x}_c|x, \mathcal{L}_c, \mathcal{E}) \tag{3}$$

Given the finite-state nature of OSTIA and GIATI models, a version of the Viterbi algorithm [14] that integrates the error-correcting decoding was developed for Eq. 2. This algorithm was applied to the source language when decoding with OSTIA transducers and to the target language in GIATI transducers, since the latter integrates smoothing techniques in the source language model.

4 Naive Bayes Bilingual Classification

In this section, an alternative approach to the joint class-conditional probability representation (Eq. 1) is presented. The idea consists in considering that the random variables s and t included in the model are independent. While this assumption is false in most real-world tasks, it often performs surprisingly well in practice. This way, Eq. 1 can be reformulated as follows:

$$\hat{c} = \operatorname*{argmax}_{c} p((s,t) \mid c) = \operatorname*{argmax}_{c} p(s \mid c) \cdot p(t \mid c) \tag{4}$$

The conditional probability $p(s \mid c)$ and $p(t \mid c)$ can be modeled in different manners. For instance, well-known monolingual text classification models can be used. Here, n-gram language models from the statistical language modelling area will be employed. An important argument that supports this decision is the existence of powerful smoothing techniques in this field.

4.1 Smooth *n*-gram Language Models

Statistical language modelling is focused on the modeling of probable sequences of linguistic units through probabilities estimated from the frequency of fixed-length sequences. In general, given a string $s = s_1 s_2 \cdots s_m$, its probability can be expressed as:

$$P(s) = \prod_{i=1}^{m} P(s_i \mid s_1 \cdots s_{i-1}) \tag{5}$$

Nevertheless, computational and parameter estimation constrains only allow for short history dependencies. Therefore, given a maximum history length n, the probability of a string s can be approximated by a n-gram language model:

$$P(s) \approx \prod_{i=1}^{m} P(s_i \mid s_{i-n+1} \cdots s_{i-1}) = \prod_{i=1}^{m} \frac{C(s_{i-n+1} \cdots s_i)}{C(s_{i-n+1} \cdots s_{i-1})} \tag{6}$$

where $C(s_i \cdots s_j)$ is the number of times that the substring $s_i \cdots s_j$ is observed in the training set.

The main problem that n-gram language models need to cope with is the estimation of probability for non-observed events (n-gram sequences) during the training process. Smoothing techniques are used to solve this issue.

These techniques, basically, consist in discounting mass probability from observed events that will be shared among non-observed events. Different discount strategies can be found in the literature, among them, Witten-Bell discount [15] in which non-observed events are modeled as seen for the first time, and modified Kneser-Ney discount [16] in which a fixed discount to all observed events is integrated in the backoff smoothing.

5 Experimental Results

5.1 Corpus

The corpus employed in these experiment was developed in the EuTrans EU project [7]. The general domain in EuTrans is that of a tourist visiting a foreign country. Specifically, in this work, the domain has been restricted to human-to-human communication in the front-desk of a hotel, which is known as the Traveller Task.

To obtain a corpus for this task, several traveller-oriented booklets were collected and those pairs of sentences fitting the above scenario were selected. To control task complexity, 16 subdomains were considered together with the pairs of sentences associated to them. The final corpus was generated from the previous "seed corpus" independently by four persons, assigning a subset of subdomains to each one. Each person defines one of the four classes A, F, J and P with a different set of subdomain coverage (see Table 1).

As it can be seen in Table 1, these four classes do overlap, therefore no perfect classification is possible and low error rates would indicate that the models employed are able to deal with the underlaying class variability.

Table 1. Subdomains in the Traveller Task and their class assignments

Classes				Subdomain	
A	F	J	P	#	Description
√	√			1	Notifying a previous reservation
√				2	Asking about rooms (availability, features, prices)
√				3	Having a look at rooms
√	√			4	Asking for rooms
√				5	Signing the registration form
√				6	Complaing about rooms
√				7	Changing rooms
	√			8	Asking for wake-up calls
	√			9	Asking for keys
	√	√		10	Asking for moving the luggage
		√		11	Notifying the departure
			√	12	Asking for the bill
		√	√	13	Asking about the bill
			√	14	Complaining about the bill
			√	15	Asking for a taxi
√	√	√	√	16	General sentences

5.2 Results

The corpus employed for the experiments consists of 8000 Spanish-English sentences extracted from the Traveller Task. Each class described above contains 2000 sentences, 1000 of them were used for training and the rest for test. The vocabulary for the Spanish sentences is 675 words and for the English sentences, 503 words.

Several experiments were carried out varying the length of the history (n-gram) of the underlaying language model. Furthermore, transducers were trained in both directions, Spanish-English (Es-En) and English-Spanish (En-Es). As far as the naive Bayes approach is concerned, two different discount techniques were used, modified Kneser-Ney discount (KN) and Witten-Bell discount (WB). The results are shown in Table 2.

The best results were obtained by applying the naive Bayes approximation using 3-gram or 4-gram language models with Witten-Bell discount. Transducer techniques achieved slightly worse results, although error rates presented by GIATI outperforms OSTIA-P in some experimental settings. The divergence on the results for GIATI depending on the source language employed, is due to the fact that the Spanish language perplexity is higher than that of the English language. An interesting open experiment would be the combination of both language pair directions to further improve the classification error rate achieved.

The difference in results between naive Bayes and transducers techniques can be explained by the better modelization of the input and output sentences in the former approach using powerful smoothing techniques, even though the unrealistic independence assumption. On the contrary, transducers generated by OSTIA-P lack of smoothing on the input language, while GIATI transducers do on the output language. Another alternative solution to the smoothing problem would be to increase the amount of training data, so that transducers would be more accurated.

In general, these results are very promising considering the complexity of the task. Indeed, when comparing the best classification error rate to monolingual

Table 2. Classification error (in %) obtained with different transducer inference and language modelling techniques. Baseline table corresponds to monolingual results using Bernoulli mixture [17] and n-gram language models

Monolingual (BASELINE)			2-grams	3-grams	4-grams
Bernoulli		Es	1.5		
n-gram	KN	Es	6.5	6.2	17.3
		En	2.5	3.2	15.3
	WB	Es	1.8	1.5	1.5
		En	1.7	1.4	1.4

Bilingual		2-grams	3-grams	4-grams
GIATI	Es-En	2.4	2.4	2.3
	En-Es	1.5	1.5	1.5
OSTIA-P	Es-En	1.7	2.3	2.5
	En-Es	2.3	2.4	2.5
Naive Bayes (n-gram)	KN	2.3	2.4	7.9
	WB	1.5	1.0	1.0

baseline cases (see Table 2), it is observed a 33% error rate reduction with respect to Bernoulli mixtures and 29% with respect to n-gram language models, when incorporating additional information (a second language) into the classification model.

6 Conclusions and Future Work

This paper has been devoted to bilingual classification endeavouring three possible approaches to attempt this task: OSTIA-P, GIATI and Naive Bayes. The idea behind these approximations was the modelization of the joint probability for a given bilingual sample (s, t). Classification error rates obtained on the Traveller Task have demonstrated the suitability of the bilingual classification models proposed, enhancing significantly previous results on the monolingual version of the same task.

As a future work remains the integration of smoothing on the input and output languages of transducers employing techniques presented in [18]. Moreover, these bilingual classification techniques are being assessed in a more general domain corpus to validate their robustness to face even more difficult tasks.

Further refinements of these models are being evaluated, considering more complex statistical models based on the combination of IBM Model 1 [13] with multinomial distributions.

References

1. McCallum, A., Nigam, K.: A comparison of event models for naive bayes text classification. In: AAAI-98 Workshop on Learning for Text Categorization. (1998)
2. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Proceedings of ECML-98, 10th European Conference on Machine Learning. Number 1398, Chemnitz, DE (1998) 137–142
3. Yang, Y.: An evaluation of statistical approaches to text categorization. Information Retrieval **1** (1999) 69–90
4. Picó, D., Casacuberta, F.: Some statistical-estimation methods for stochastic finite-state transducers. Machine Learning **44** (2001) 121–142
5. Knight, K., Al-Onaizan, Y.: Translation with finite-state devices. In: Third Conference of the Association for Machine Translation in the Americas. Volume 1529., Langhorne, PA, USA (1998) 421–437
6. Vidal, E.: Finite-state speech-to-speech translation. In: Int. Conf. on Acoustics Speech and Signal Processing, Vol.1, Munich, Germany (1997) 111–114
7. Amengual, J.C., Benedí, J.M., Castano, A., Castellanos, A., Jiménez, V.M., Llorens, D., Marzal, A., Pastor, M., Prat, F., Vidal, E., Vilar, J.M.: The EuTrans-I speech translation system. Machine Translation **15** (2000) 75–103
8. Oncina, J., García, P., Vidal, E.: Learning subsequential transducers for pattern recognition interpretation tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence **15** (1993) 448–458
9. Gold, E.M.: Language identification in the limit. Information and Control **10** (1967) 447–474
10. Oncina, J., Varó, M.A.: Using domain information during the learning of a subsequential transducer. In: ICGI, Berlin, Germany (1996) 301–312
11. Cubel, E.: Aprendizaje de transductores subsecuenciales estocásticos. Technical Report II-DSIC-B-23/01, Universidad Politécnica de Valencia, Spain (2002)
12. Och, F.J., Ney, H.: Improved statistical alignment models. In: ACL00, Hong Kong, China (2000) 440–447
13. Brown, P.F., Pietra, S.D., Pietra, V.J.D., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics **19** (1993) 263–312
14. Viterbi, A.: Error bounds for convolutional codes and a asymtotically optimal decoding algorithm. IEEE Transactions on Information Theory **13** (1967) 260–269
15. Witten, I.H., Bell, T.C.: The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. IEEE Trans. Information Theory **37** (1991) 1085–1094
16. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modelling. In: Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, San Francisco, USA (1996) 310–318
17. Juan, A., Vidal, E.: On the use of bernoulli mixture models for text classification. In: Workshop on Pattern Recognition in Information Systems (PRIS 01), Setúbal (Portugal) (2001)
18. Llorens, D.: Suavizado de autómatas y traductores finitos estocásticos. PhD thesis, Universitat Politècnica de València (2000) Advisor(s): Dr. J. M. Vilar and Dr. F. Casacuberta.

Semantic Similarity Between Sentences Through Approximate Tree Matching

Francisco Jose Ribadas, Manuel Vilares, and Jesus Vilares

[1] Computer Science Dept., Univ. of Vigo
Campus As Lagoas, s/n, 32004 Ourense, Spain
{ribadas,vilares}@uvigo.es
[2] Computer Science Dept., Univ. of A Coruña
Campus de Elviña, s/n, 15071 A Coruña, Spain
jvilares@udc.es

Abstract. We describe an algorithm to measure the similarity between sentences, integrating the edit distance between trees and single-term similarity techniques, and also allowing the pattern to be defined approximately, omitting some structural details. A technique of this kind is of interest in a variety of applications, such as information extraction/retrieval or question answering, where error-tolerant recognition allows incomplete sentences to be integrated in the computation process.

1 Introduction

Many different approaches have been applied on computing similarity between documents. Some of them consider a vector space model or semantically-based word-word proximity [3]; others incorporate syntactic [2] or structural information, which can be dealt with by pattern-matching. At this point, document similarity based on matching sentences seems to have a more significant effect on the quality and effectiveness of the resulting measures, due to its robustness in dealing with noisy terms.

So, finding sentence similarity models is a central question in dealing with document similarity. In this context, most efforts have been targeted toward single-terms based techniques [1]. The work on the combination of parse structures and single-terms is limited, and there is nothing at all related to dealing with incomplete or vague structures. Our proposal seeks to provide these capabilities in a new sentence similarity measure, integrating semantic similarity with *variable length don't care* matching on parse trees.

2 The Tree Edit Distance

Given a pattern tree, P and a data tree, D, we define an *edit operation* as a pair $a \to b$, $a \in labels(P) \cup \{\varepsilon\}$, $b \in labels(D) \cup \{\varepsilon\}$, where ε represents the empty string and $labels(T)$ the set of symbols labeling nodes in tree T. We can delete $(a \to \varepsilon)$, insert $(\varepsilon \to b)$, and change a node $(a \to b)$. Each edit operation has

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 638–646, 2005.

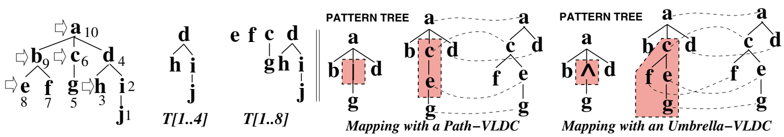


Fig. 1. Inverse postorder and VLDC symbols.

a cost, $\gamma(a \rightarrow b)$, which satisfies the properties of a metric. The *edit distance* between two ordered trees P and D, $\delta(P, D)$, is defined as the cost of the sequence of edit operations that transform one tree into the other with a minimal cost.

We compute a modified edit distance, following Zhang *et al.* [5], where data trees can be simplified by removing some irrelevant subtrees with no cost. Given an inverse postorder traversal to name each node i of tree T by $t[i]$, we introduce $r(i)$ as the rightmost leaf descendant of the subtree rooted at $t[i]$, and $T[i..j]$ as the ordered subforest of T induced by the nodes numbered i to j, as shown Fig. 1. In particular, we have $T[r(i)..i]$ as the tree rooted at $t[i]$, denoted as $T[i]$. We also introduce *r_keyroots*(T) as the set of all nodes in T which have right siblings plus the root, *root*(T), of T, indicated by arrows in Fig. 1.

The value we want to compute is the edit distance between the pattern tree and the data tree resulting from an optimal set of cost-free cuts that yields a minimal distance. Cutting at node $d[j]$ means removing the subtree $D[j]$ from the data tree D. Formally, given trees P and D, the *forest edit distance with cuts* between a target subforest $P[s_1..s_2]$ and a data subforest $D[t_1..t_2]$, is defined as:

$$fd(s_1..s_2, t_1..t_2) = \; min_{S \in subtrees(D[t_1..t_2])}\{\delta(P[s_1..s_2], cut(D[t_1..t_2], S))\}$$

where $cut(D[t_1..t_2], \mathcal{S})$ is the subforest $D[t_1..t_2]$ with subtree removals at all nodes included in the set $\mathcal{S}$, and where $subtrees(D[t_1..t_2])$ is the set of all possible sets of subtree cuts in $D[t_1..t_2]$. We will use the notation $td(s_2, t_2)$ when the two subforests are composed by only one subtree, that is, $s_1 = r(s_2)$ and $t_1 = r(t_2)$.

Zhang *et al.* propose a dynamic programming algorithm to compute this tree edit distance, $td(P, D)$, in a bottom-up fashion, first determining distances from all leaf r_keyroots, then for r_keyroots at the next level, and so on to the root.

We also support the use of *variable length don't care* (VLDC) symbols in the pattern tree, which allows us to omit structural details and manage more general patterns. We consider two definitions for VLDC matching [5], shown in Fig. 1:

- The VLDC substitutes part of a path from the root to a leaf of the data tree. We represent this with "|" and call it a *path*-VLDC.
- The VLDC matches part of such a path and the subtrees emanating from nodes of that path. We call this an *umbrella*-VLDC, represented with "$\wedge$".

To formalize the use of VLDC requires to introduce the notion of VLDC-*substitution*, that assigns to each VLDC node in a pattern tree P a set of nodes taken from the data tree D. The final edit distance is computed between the tree $\bar{P}$, resulting from an optimal VLDC-substitution on P, and the data tree D.

To compute distances involving $\wedge$-VLDC symbols it is necessary to use an auxiliary distance. The *suffix forest distance* between F_P and F_D, forests in trees P and D, denoted $sfd(F_P, F_D)$, is the distance between F_P and $\bar{F}_D$, where $\bar{F}_D$ is a subforest of F_D with some consecutive subtrees, all having the same parent, removed from the right. Formally, we have $sfd(F_P, F_D) = \min_{\bar{F}_D}\{fd(F_P, \bar{F}_D)\}$.

3 Semantic Similarity

We want to extend the matching algorithm in a similarity measure for parse trees taking into account the semantic proximity between words. In essence, we will propagate a similarity measure at word level through the nodes in accordance with the syntactic distances computed by tree matching.

3.1 Semantic Similarity at Word Level

We follow Lin's work [1], based on the WordNet taxonomy, a computer dictionary that implements a set of semantic relationships between pairs of structures called *synsets*, that represent sets of words sharing the same meaning.

Lin uses the hyperonymy relationship, i. e. the classical *is-a* relation between a general concept and another more specific one, and his approach is based on the information content of the synsets. Given two words, w_1 and w_2, belonging to the synsets S_1 and S_2, respectively, let $P(S_i)$ be the probability that a randomly selected word from WordNet belongs to synset S_i or one of its more specific synsets. If S_0 is the most specific synset that subsumes both S_1 and S_2 according to the hyperonymy relation, the similarity between w_1 and w_2 is:

$$sim_{\text{Lin}}(w_1, w_2) = \frac{2 \times logP(S_0)}{logP(S_1)\ +\ logP(S_2)}$$

We use Lin's measure as a basis to compute the semantic cost, γ_{sem}, associated with the edit operations applied to the words in the sentences we are comparing. Being w_i and w_j two words or ε, we compute $\gamma_{\text{sem}}(w_i \to w_j)$, as follows:

$$\gamma_{\text{sem}}(w_i \to w_j) = \begin{cases} 1 & \textbf{if } w_i = \varepsilon \textbf{ or } w_j = \varepsilon \\ 1 - sim_{\text{Lin}}(w_i, w_j) & \textbf{if } w_i \neq \varepsilon \textbf{ and } w_j \neq \varepsilon \end{cases}$$

3.2 Sentence Similarity Based on Edit Distance

Our aim is to compute a semantic distance between pairs of whole sentences, taking into account both lexical and syntactic information. As a first approach we could use a classical string matching algorithm [4] to locate the operations to be applied over the words in the sentences. The algorithm will identify the sequence of modifications needed to obtain one of them from the other with a minimum cost. We will simply treat each word in the sentences as if it were a character, measuring the cost of the edit operations with the γ_{sem} function.

As shown in Fig. 2, the correspondence between words that this approach yields can be rather odd. This method aims to get a minimum cost alignment, without taking into account whether it makes sense or not. So, in the first pair

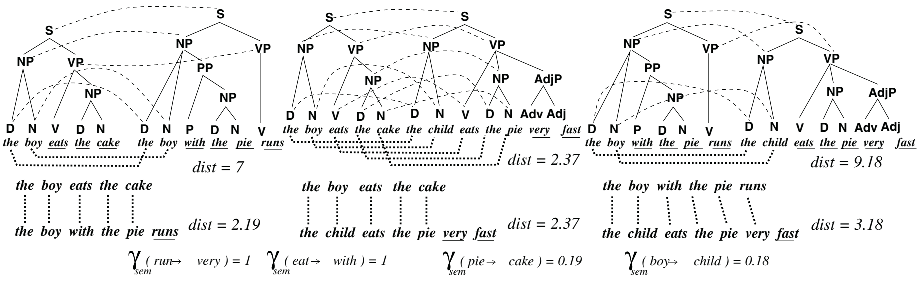

the boy eats the cake
the boy with the pie runs
dist = 7
dist = 2.19
the boy eats the cake
the child eats the pie very fast
dist = 2.37
dist = 2.37
the boy with the pie runs
the child eats the pie very fast
dist = 9.18
dist = 3.18
γ_sem (run→ very) = 1
γ_sem (eat→ with) = 1
γ_sem (pie→ cake) = 0.19
γ_sem (boy→ child) = 0.18

Fig. 2. String matching and tree matching mapping between words.

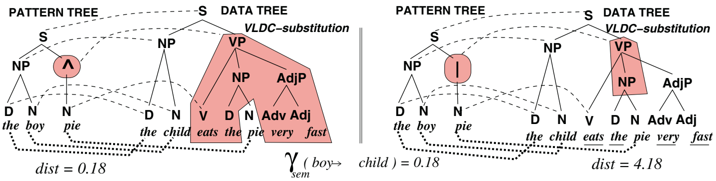


Fig. 3. Semantic distance with VLDC symbols.

of sentences in Fig. 2, the algorithm aligns the verb "*eat*" with the preposition "*with*", which has no grammatical sense since their syntactic roles are different.

This behaviour masks the actual semantic proximity and justifies to consider tree matching to identify the transformations from which semantic proximity should be computed. The sets of partial distances between subforests computed by Zhang *et al.* [5] algorithm are used to get a minimal cost mapping, shown with dashed lines in Fig. 2. From this, we extract the edit operations applied over leaves which give us the set of words whose semantic costs, computed by γ_{sem}, are accumulated.

4 Semantic Similarity from Word to Sentence Level

Having outlined our proposal, we show how the propagation of semantic measures is performed. Semantic and syntactic distances are computed in parallel, guided by the computation of partial subforest distances.

The *semantic forest distance*, fd_{sem}, is an extension of the concept at syntactic level [5]. Given trees P and D, $fd_{\text{sem}}(s_1..s_2, t_1..t_2)$ is the semantic cost of the edit operations performed on the words belonging to the subforests $P[s_1..s_2]$ and $D[t_1..t_2]$. The selection of these operations is made according to the actual edit operation applied on the associated leaves to obtain the optimal $fd(s_1..s_2, t_1..t_2)$ value. In an analogous way, we have $td_{\text{sem}}(i, j) = fd_{\text{sem}}(r(i)..i, r(j)..j)$.

4.1 Computation of Semantic Distances Without VLDC

What our approach does is to compute partial distances between the syntactic structure of the two parse trees, and use these fd values to decide which semantic values, fd_{sem}, should be propagated in each case. When no VLDC symbols are involved this propagation is made according to the following set of formulae, where nodes $s \in P[i]$, $t \in D[j]$, and $i \in r\_keyroots(P)$ and $j \in r\_keyroots(D)$:

$$
\begin{array}{l}
(1)\ fd_{\text{sem}}(\emptyset,\emptyset) = 0 \\
\quad fd_{\text{sem}}(\emptyset, r(j)..t) = \begin{cases} fd_{\text{sem}}(\emptyset, r(j)..t-1) + \gamma_{\text{sem}}(\varepsilon \rightarrow word(d[t])) & \textbf{if } r(t) = t \\ fd_{\text{sem}}(\emptyset, r(j)..t-1) & \textbf{otherwise} \end{cases} \\
\quad fd_{\text{sem}}(r(i)..s, \emptyset) = \begin{cases} fd_{\text{sem}}(r(i)..s-1, \emptyset) + \gamma_{\text{sem}}(word(p[s]) \rightarrow \varepsilon) & \textbf{if } r(s) = s \\ fd_{\text{sem}}(r(i)..s-1, \emptyset) & \textbf{otherwise} \end{cases}
\end{array}
$$

$$
(2)\ fd_{\text{sem}}(r(i)..s, r(j)..t) = min \left\{ \begin{array}{l}
\left\{ \begin{array}{ll} fd_{\text{sem}}(r(i)..s-1, r(j)..t) + \gamma_{\text{sem}}(word(p[s]) \rightarrow \varepsilon) & \textbf{if } r(s) = s \\ fd_{\text{sem}}(r(i)..s-1, r(j)..t) & \textbf{otherwise} \end{array} \right\} \\
\qquad \textbf{if } fd(r(i)..s, r(j)..t) = fd(r(i)..s-1, r(j)..t) + \gamma(p[s] \rightarrow \varepsilon) \\
\left\{ \begin{array}{ll} fd_{\text{sem}}(r(i)..s, r(j)..t-1) + \gamma_{\text{sem}}(\varepsilon \rightarrow word(d[t])) & \textbf{if } r(t) = t \\ fd_{\text{sem}}(r(i)..s, r(j)..t-1) & \textbf{otherwise} \end{array} \right\} \\
\qquad \textbf{if } fd(r(i)..s, r(j)..t) = fd(r(i)..s, r(j)..t-1) + \gamma(\varepsilon \rightarrow d[t]) \\
\left\{ \begin{array}{ll} fd_{\text{sem}}(r(i)..s-1, r(j)..t-1) + & \\ \qquad \gamma_{\text{sem}}(word(p[s]) \rightarrow word(d[t])) & \textbf{if } r(s) = s \ \textbf{and } r(t) = t \\ fd_{\text{sem}}(r(i)..s-1, r(j)..t-1) & \textbf{otherwise} \end{array} \right\} \\
\qquad \textbf{if } fd(r(i)..s, r(j)..t) = fd(r(i)..s-1, r(j)..t-1) + \gamma(p[s] \rightarrow d[t]) \\
fd_{\text{sem}}(r(i)..s, \emptyset) + \sum_{d[k] \in \text{leaves}(D[t])} \gamma_{\text{sem}}(\varepsilon \rightarrow word(d[k])) \\
\qquad \textbf{if } fd(r(i)..s, r(j)..t) = fd(r(i)..s-1, \emptyset)
\end{array} \right.
$$

$$
\textbf{if } r(s) = r(i) \textbf{ and } r(t) = r(j) \textbf{ and } p[s] \neq "\wedge" \textbf{ and } p[s] \neq "\mid"
$$

$$
(3)\ fd_{\text{sem}}(r(i)..s, r(j)..t) = min \left\{ \begin{array}{l}
\left\{ \begin{array}{ll} fd_{\text{sem}}(r(i)..s-1, r(j)..t) + \gamma_{\text{sem}}(word(p[s]) \rightarrow \varepsilon) & \textbf{if } r(s) = s \\ fd_{\text{sem}}(r(i)..s-1, r(j)..t) & \textbf{otherwise} \end{array} \right\} \\
\qquad \textbf{if } fd(r(i)..s, r(j)..t) = fd(r(i)..s-1, r(j)..t) + \gamma(p[s] \rightarrow \varepsilon) \\
\left\{ \begin{array}{ll} fd_{\text{sem}}(r(i)..s, r(j)..t-1) + \gamma_{\text{sem}}(\varepsilon \rightarrow word(d[t])) & \textbf{if } r(t) = t \\ fd_{\text{sem}}(r(i)..s, r(j)..t-1) & \textbf{otherwise} \end{array} \right\} \\
\qquad \textbf{if } fd(r(i)..s, r(j)..t) = fd(r(i)..s, r(j)..t-1) + \gamma(\varepsilon \rightarrow d[t]) \\
fd_{\text{sem}}(r(i)..r(s)-1, r(j)..r(t)-1) + td_{\text{sem}}(s,t) \\
\qquad \textbf{if } fd(r(i)..s, r(j)..t) = fd(r(i)..r(s)-1, r(j)..r(t)-1) + td(s,t) \\
fd_{\text{sem}}(r(i)..s, r(j)..r(t)-1) + \sum_{d[k] \in \text{leaves}(D[t])} \gamma_{\text{sem}}(\varepsilon \rightarrow word(d[k])) \\
\qquad \textbf{if } fd(r(i)..s, r(j)..t) = fd(r(i)..s-1, r(t)-1)
\end{array} \right.
$$

$$
\textbf{otherwise}
$$

We have three formulae (initialization, subtree-to-subtree and subforest-to-subforest distances) that reflect how is identified the operation from which the current $fd(r(i)..s, r(j)..t)$ was obtained, and how new fd_{sem} values are computed.

Zhang *et al.* [5] compute bottom-up partial distances, increasing the sizes of the subforests being considered. In each step, every way of obtaining the current distance, one for each edit operation and one for subtree cuts, is computed and the minimum value is recorded. In our proposal, for each syntactic distance we identify the previous partial ones used to compute it, and use them to determine which edit operation was applied. Once the best edit operation has been identified, we compute the current $fd_{\text{sem}}(r(i)..s, r(j)..t)$ value. If no leaves were involved, we simply propagate the semantic distance associated with the previous fd value from which $fd(r(i)..s, r(j)..t)$ was obtained. Otherwise, when $r(s) = s$ or $r(t) = t$, we add to this fd_{sem} value the semantic cost of the operation to be applied on those leaves.

In the case of subtree cuts, previous values are incremented with the cost of deleting every word in the cut subtrees, as shown in the last cases of formulae (2) and (3). So, although in syntactic matching some parts of the data trees can be omitted, our proposal does not miss out the semantic cost of the deleted words.

4.2 Computation of Semantic Distances with VLDC

When cost-free cuts are allowed, the two types of VLDC's can be interchanged without affecting the final syntactic distance [5]. However, it is possible to differentiate their interpretation when semantic distances are computed.

S → NP VP
NP → Name
NP → Det Name
NP → Det Name PP
NP → Det Name PP PP
PP → Prep NP
AdjP → Adv Adj
VP → Verb NP AdjP
VP → Verb NP PP AdjP

$$\text{"the} \left\{\begin{array}{c}\text{boy}\\ \text{child}\\ \text{girl}\end{array}\right\} \text{with the} \left\{\begin{array}{c}\text{cake}\\ \text{pie}\end{array}\right\} \text{of} \left[\text{the} \left\{\begin{array}{c}\text{friend}\\ \text{mate}\end{array}\right\} \text{of}\right]^i \text{Mary} \left\{\begin{array}{c}\text{runs}\\ \text{walks}\end{array}\right\} \text{very fast"}$$

$$\text{"the} \left\{\begin{array}{c}\text{boy}\\ \text{child}\\ \text{girl}\end{array}\right\} \left\{\begin{array}{c}\text{eats}\\ \text{steals}\end{array}\right\} \text{the} \left\{\begin{array}{c}\text{cake}\\ \text{pie}\end{array}\right\} \text{of} \left[\text{the} \left\{\begin{array}{c}\text{friend}\\ \text{mate}\end{array}\right\} \text{of}\right]^i \text{Mary very fast"}$$

Fig. 4. Toy English grammar and data sentences.

For | -VLDC we want the words in the cut leaves, underlined in the rightmost side of Fig. 3, to be taken into account in the semantic distance. Whereas, with the ∧-VLDC, words in subtrees not taken into account in the syntactic distance will not contribute to the final semantic distance. These subtrees, shaded in Fig. 3, are identified when the auxiliary values sfd are computed.

We shall start with the formulae for tree-to-tree distances with | -VLDC. Being $p[s] = " \mid "$, $r(i) = r(s)$, $r(j) = r(t)$ and nodes $d[t_k]$, $1 \leq k \leq n_t$, children of $d[t]$:

$$fd_{\text{sem}}(r(i)..s, r(j)..t) = min \left\{ \begin{array}{l}
\left\{ \begin{array}{ll} fd_{\text{sem}}(r(i)..s-1, r(j)..t) + \gamma_{\text{sem}}(word(p[s]) \to \varepsilon) & \textbf{if } r(s) = s \\ fd_{\text{sem}}(r(i)..s-1, r(j)..t) & \textbf{otherwise} \end{array} \right\} \\
\qquad \textbf{if } fd(r(i)..s, r(j)..t) = fd(r(i)..s-1, r(j)..t) + \gamma(p[s] \to \varepsilon) \\
\left\{ \begin{array}{ll} fd_{\text{sem}}(r(i)..s, r(j)..t-1) + \gamma_{\text{sem}}(\varepsilon \to word(d[t])) & \textbf{if } r(t) = t \\ fd_{\text{sem}}(r(i)..s, r(j)..t-1) & \textbf{otherwise} \end{array} \right\} \\
\qquad \textbf{if } fd(r(i)..s, r(j)..t) = fd(r(i)..s, r(j)..t-1) + \gamma(\varepsilon \to d[t]) \\
\left\{ \begin{array}{ll} fd_{\text{sem}}(r(i)..s-1, r(j)..t-1) + & \\ \qquad \gamma_{\text{sem}}(word(p[s]) \to word(d[t])) & \textbf{if } r(s) = s \textbf{ and } r(t) = t \\ fd_{\text{sem}}(r(i)..s-1, r(j)..t-1) & \textbf{otherwise} \end{array} \right\} \\
\qquad \textbf{if } fd(r(i)..s, r(j)..t) = fd(r(i)..s-1, r(j)..t-1) + \gamma(p[s] \to d[t]) \\
fd_{\text{sem}}(\emptyset, r(j)..t-1) + td_{\text{sem}}(s, t_k) - td_{\text{sem}}(\emptyset, t_k) \\
\qquad \textbf{if } fd(r(i)..s, r(j)..t) = fd(\emptyset, r(j)..t-1) + td(s, t_k) - td(\emptyset, t_k) \\
\qquad\qquad \text{where } t_k = argmin_{1 \leq k \leq n_t}\{td(s, t_k) - td(\emptyset, t_k)\} \\
fd_{\text{sem}}(r(i)..s, \emptyset) + \sum_{d[k] \in \text{leaves}(D[t])} \gamma_{\text{sem}}(\varepsilon \to word(d[k])) \\
\qquad \textbf{if } fd(r(i)..s, r(j)..t) = fd(r(i)..s-1, \emptyset)
\end{array} \right.$$

To manage ∧-VLDC as we outlined above, being $d[t_k]$, $1 \leq k \leq n_t$ the children of $d[t]$, when $r(i) = r(s)$, $r(j) = r(t)$ and $p[s] = " \wedge "$ we have that:

$$fd_{\text{sem}}(r(i)..s, r(j)..t) = min \left\{ \begin{array}{l}
\left\{ \begin{array}{ll} fd_{\text{sem}}(r(i)..s-1, r(j)..t) + \gamma_{\text{sem}}(word(p[s]) \to \varepsilon) & \textbf{if } r(s) = s \\ fd_{\text{sem}}(r(i)..s-1, r(j)..t) & \textbf{otherwise} \end{array} \right\} \\
\qquad \textbf{if } fd(r(i)..s, r(j)..t) = fd(r(i)..s-1, r(j)..t) + \gamma(p[s] \to \varepsilon) \\
\left\{ \begin{array}{ll} fd_{\text{sem}}(r(i)..s, r(j)..t-1) + \gamma_{\text{sem}}(\varepsilon \to word(d[t])) & \textbf{if } r(t) = t \\ fd_{\text{sem}}(r(i)..s, r(j)..t-1) & \textbf{otherwise} \end{array} \right\} \\
\qquad \textbf{if } fd(r(i)..s, r(j)..t) = fd(r(i)..s, r(j)..t-1) + \gamma(\varepsilon \to d[t]) \\
\left\{ \begin{array}{ll} fd_{\text{sem}}(r(i)..s-1, r(j)..t-1) + & \\ \qquad \gamma_{\text{sem}}(word(p[s]) \to word(d[t])) & \textbf{if } r(s) = s \textbf{ and } r(t) = t \\ fd_{\text{sem}}(r(i)..s-1, r(j)..t-1) & \textbf{otherwise} \end{array} \right\} \\
\qquad \textbf{if } fd(r(i)..s, r(j)..t) = fd(r(i)..s-1, r(j)..t-1) + \gamma(p[s] \to d[t]) \\
td_{\text{sem}}(s, t_k) \\
\qquad \textbf{if } fd(r(i)..s, r(j)..t) = td(s, t_k) \\
\qquad\qquad \text{where } t_k = argmin_{1 \leq k \leq n_t}\{td(s, t_k)\} \\
sfd_{\text{sem}}(r(i)..s-1, r(j)..t_k) \\
\qquad \textbf{if } fd(r(i)..s, r(j)..t) = sfd(r(i)..s-1, r(j)..t_k) \\
\qquad\qquad \text{where } t_k = argmin_{1 \leq k \leq n_t}\{sfd(r(i)..s-1, r(j)..t_k)\}
\end{array} \right.$$

To ensure the desired semantic behaviour for ∧-VLDC, we introduce the *semantic suffix forest distance* and specify how these values are computed. Being F_P and F_D, forests in the tree P and the data tree D, the *semantic suffix forest distance*, $sfd_{\text{sem}}(F_P, F_D)$, is the semantic distance between F_P and $\bar{F}_D$, where $\bar{F}_D$ is a subforest of F_D with some consecutive subtrees, all having the same parent, removed from the right. Formally, $sfd_{\text{sem}}(F_P, F_D) = \min_{\bar{F}_D}\{fd_{\text{sem}}(F_P, \bar{F}_D)\}$.

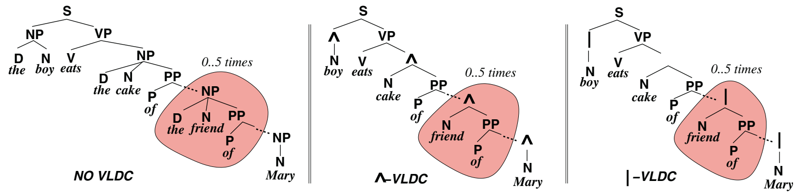


Fig. 5. Pattern trees.

As result, we get the required behaviour for ∧-VLDC, since words present in removed subtrees will not be taken into account in the final semantic distance. The computation procedure propagates semantic values in the following way:

$$(1)\ sfd_{\text{sem}}(\emptyset,\emptyset) = 0$$

$$sfd_{\text{sem}}(\emptyset, r(j)..t) = \begin{cases} 0 & \textbf{if } r(t)=r(j) \textbf{ or } r(parent(t))=r(j) \\ \left.\begin{cases} sfd_{\text{sem}}(\emptyset, r(j)..t-1) + \gamma_{\text{sem}}(\varepsilon \to word(d[t])) & \textbf{if } r(t)=t \\ sfd_{\text{sem}}(\emptyset, r(j)..t-1) & \textbf{otherwise} \end{cases}\right\} & \textbf{otherwise} \end{cases}$$

$$sfd_{\text{sem}}(r(i)..s,\emptyset) = fd_{\text{sem}}(r(i)..s,\emptyset)$$

$$(2)\ sfd_{\text{sem}}(r(i)..s, r(j)..t) = \begin{cases} fd_{\text{sem}}(r(i)..s,\emptyset) & \textbf{if } fd(r(i)..s,\emptyset) < fd(r(i)..s, r(j)..t) \\ fd_{\text{sem}}(r(i)..s, r(j)..t) & \textbf{otherwise} \end{cases}$$

$$\textbf{if } r(t) = r(j)$$

$$(3)\ sfd_{\text{sem}}(r(i)..s, r(j)..t) = min \begin{cases} \left.\begin{cases} sfd_{\text{sem}}(r(i)..s-1, r(j)..t) + \gamma_{\text{sem}}(word(p[s]) \to \varepsilon) & \textbf{if } r(s)=s \\ sfd_{\text{sem}}(r(i)..s-1, r(j)..t) & \textbf{otherwise} \end{cases}\right\} \\ \qquad \textbf{if } sfd(r(i)..s, r(j)..t) = sfd(r(i)..s-1, r(j)..t) + \gamma(p[s] \to \varepsilon) \\ \left.\begin{cases} sfd_{\text{sem}}(r(i)..s, r(j)..t-1) + \gamma_{\text{sem}}(\varepsilon \to word(d[t])) & \textbf{if } r(t)=t \\ sfd_{\text{sem}}(r(i)..s, r(j)..t-1) & \textbf{otherwise} \end{cases}\right\} \\ \qquad \textbf{if } sfd(r(i)..s, r(j)..t) = sfd(r(i)..s, r(j)..t-1) + \gamma(\varepsilon \to d[t]) \\ sfd_{\text{sem}}(r(i)..r(s)-1, r(j)..r(t)-1) + td_{\text{sem}}(s,t) \\ \qquad \textbf{if } sfd(r(i)..s, r(j)..t) = sfd(r(i)..r(s)-1, r(j)..r(t)-1) + td(s,t) \end{cases}$$

$$\textbf{otherwise}$$

By means of these formulae we get two kinds of semantic matching: a restricted match with | -VLDC, where precise patterns can be easily described; and a more flexible and vague method, using ∧-VLDC, able to match a wider range of trees.

5 A Practical Example

To illustrate our proposal we have created a bank of data sentences following the grammar and the scheme shown in Fig. 4, where i, with values from 0 to 7, is the number of repetitions of the substrings enclosed in braces. This evaluation frame provides a highly ambiguous evaluation environment, with a number of parses growing exponentially with i. In the case of the tree matching based approach three sets of pattern trees were built from the deterministic parse for sentences of the form *"the boy eats the cake of* [*the friend of*]^i, shown in Fig. 5. To test string matching, pattern and data sentences were used without parsing them.

The criteria used to compare these methods is the number of sentences whose similarity values with respect to a pattern sentence, normalized in $[0, 1]$, fall under a given threshold, assuming that more precise approaches should provide reduced sets of sentences. We aim to determine whether the obtained values are able to make the semantic proximity between sentences evident, and to identify non-similar ones without deviations.

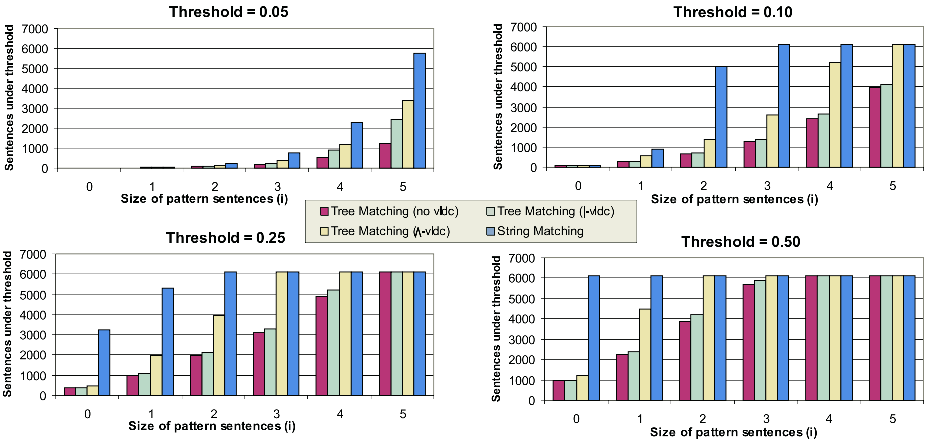


Fig. 6. Tree matching vs. string matching semantic measures.

Practical results are shown in Fig. 6, considering thresholds 0.05, 0.10, 0.25 and 0.50. Under the four thresholds our proposal seems to be more precise, showing a higher discrimination power and returning a smaller number of sentences. The string based approach suffers deviations when sentences being compared share the same words with different syntactic roles, as is the case here. So, the number of sentences under all thresholds is always high, showing that tree based values provide a more accurate notion of semantic proximity.

With regard to the VLDC matching, it offers intermediate results, closer to non-VLDC tree matching. As expected, the | -VLDC gives more accurate measures, meanwhile the ∧-VLDC patterns are slightly less precise, but without suffering the rapid degeneration shown by the string based measure for highest thresholds. In these cases, this method identifies all of the data sentences as being similar to the given pattern, without making any distinction between them, since the string based metric assigns a high relevance to word-to-word correspondences, making other syntactic and semantic aspects irrelevant.

6 Conclusions

We exploit the meaning of single-terms by integrating it into the edit distance computation, allowing to take advantage of the use of semantic information in pattern-matching processes. Preliminary results seem to support our approach opposed to solutions based exclusively on syntactic structures or single terms.

References

1. D. Lin. An information-theoretic definition of similarity. *Proc. of 15th Int. Conf. on Machine Learning*, 296–304, 1998.

2. Smeaton, A.F, O'Donell, R., Kelley,F.: Indexing Structures Derived from Syntax in TREC-3: System Description. *Proc. of* 3^{rd} *Text REtrieval Conference*, 1994.
3. A.F. Smeaton, I. Quigley. Experiments on using semantic distances between words in image caption retrieval. *Proc. of the 19th Annual Int. ACM Conf. on Research and Development in Information Retrieval*, 174–180, 1996.
4. R.A. Wagner and M.J. Fischer. The string to string correction problem. *Journal of the ACM*, 21(1):168–173, 1974.
5. K. Zhang, D. Shasha, and J.T.L. Wang. Approximate tree matching in the presence of variable length don't cares. *Journal of Algorithms*, 16(1):33–66, 1994.

Part X

Applications

A Text Categorization Approach for Music Style Recognition

Carlos Pérez-Sancho, José M. Iñesta, and Jorge Calera-Rubio

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante, E-03080 Alicante, Spain
{cperez,inesta,calera}@dlsi.ua.es
http://grfia.dlsi.ua.es

Abstract. The automatic classification of music files into styles is one challenging problem in music information retrieval and for music style perception understanding. It has a number of applications, like the indexation and exploration of musical databases. Some techniques used in text classification can be applied to this problem. The key point is to establish a music equivalent to the words in texts. A number of works use the combination of intervals and duration ratios for music description. In this paper, different statistical text recognition algorithms are applied to style recognition using this kind of melody representation, exploring their performance for different word sizes and statistical models.

1 Introduction

Machine learning and pattern recognition techniques, successfully employed in other fields, can be also applied to music analysis. One of the tasks that can be posed is the modelization of the music style. Immediate applications are the classification, indexation, and content-based search in digital music libraries, where digitised (MP3), sequenced (MIDI) or structurally represented (XML) music can be found. Some recent papers explore the capabilities of these methods to recognise music style, either using audio [1–3], or symbolic sources [4–6].

Our aim is to explore the capabilities of text categorization algorithms to solve problems relevant to computer music. In this paper, some of those methods are applied to the recognition of musical genres from a symbolic representation of the melody. Some styles like jazz, classical, ragtime or gregorian have been chosen as an initial benchmark for the proposed methodology due to the general agreement in the musicology community about their definition and limits.

2 Methodology

2.1 Data Sets

Experiments were performed using two different corpora. Both of them are made up of MIDI files, containing monophonic sequences.

The first corpus is a set of MIDI files from *Jazz* and *Classical* music collected from different web sources, without any processing before entering the system.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 649–657, 2005.

The melodies are real-time sequenced by musicians, without quantization. The corpus is made up of 110 MIDI files, 45 of them being classical music and 65 being jazz music. The length of the corpus is around 10,000 bars (40,000 beats).

The second corpus is that used by Cruz et al. [4]. It consists of 300 MIDI files, from three different styles: *Gregorian*, passages from the sacred music of *J. S. Bach* (Baroque), and *Scott Joplin* ragtimes; with 100 files per class. In this corpus, the melodies are step-by-step sequenced, and much shorter (around 10 bars per file in average) than in the former case.

2.2 Encoding

The encoding used in this work has been inspired in the encoding method proposed in [7]. This method generates n-grams of notes from the melodies, encoding pitch interval information and a kind of duration ratios. The encoding of MIDI files is performed as follows:

Melody Extraction. The melody track from each MIDI file is analyzed, obtaining the pitch and duration for each note. Durations are calculated as the number of ticks between the onset of the note and that of the next, ingnoring all intermediate rests. As a result, a pair of values $\{pitch, duration\}$ is obtained for each note in the analyzed track.

n-Word Extraction. Next, the melody is divided into n-note windows. For each window, a sequence of intervals and duration ratios is obtained (see Fig. 1 for an example), calculated using Eqs. (1) and (2) respectively.

$$I_i = Pitch_{i+1} - Pitch_i \qquad (i = 1, \ldots, n-1) \qquad (1)$$

$$R_i = \frac{Onset_{i+2} - Onset_{i+1}}{Onset_{i+1} - Onset_i} \qquad (i = 1, \ldots, n-1) \qquad (2)$$

Each n-word is defined then as a sequence of symbols: $[I_1 \; R_1 \; \ldots \; I_{n-1} \; R_{n-1}]$.

n-Words Coding. The n-words obtained in the previous step are mapped into sequences of alphanumeric characters, which we will name n-words due to the equivalence we want to establish between melody and text. As the number of ASCII printable characters is lower than all possible intervals, a non-linear mapping is used to assign characters to different interval values (see [7] for details).

Stop Words. A melody can contain pairs of notes separated by long rests, that can last for some seconds, or even minutes. Consecutive notes separated by this kind of rests are not really related, so the next note can be considered as the beginning of a new melody. Therefore, it is not fair encoding together consecutive notes separated by a large rest in a melody.

Considering this, a silence threshold is established, in a way that when a rest longer than this threshold is found, no words are generated across it. This restriction implies that, for each rest longer than this threshold, $n-1$ words less are encoded. This threshold has been empirically set to a rest of four beats.

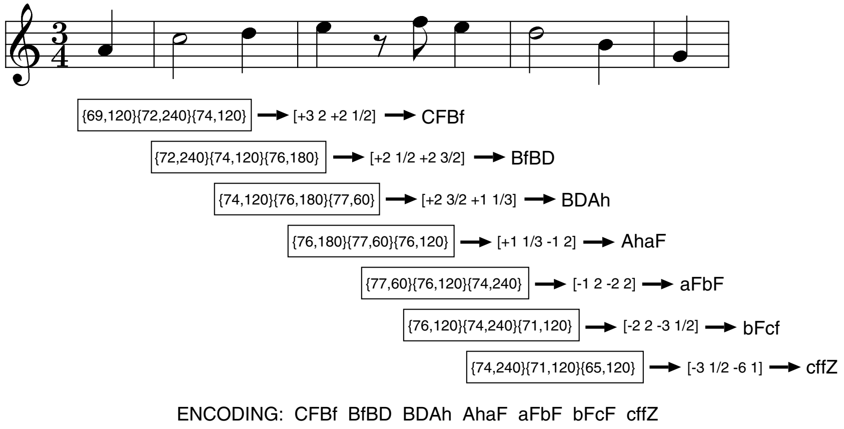

{69,120}{72,240}{74,120} → [+3 2 +2 1/2] → CFBf
{72,240}{74,120}{76,180} → [+2 1/2 +2 3/2] → BfBD
{74,120}{76,180}{77,60} → [+2 3/2 +1 1/3] → BDAh
{76,180}{77,60}{76,120} → [+1 1/3 -1 2] → AhaF
{77,60}{76,120}{74,240} → [-1 2 -2 2] → aFbF
{76,120}{74,240}{71,120} → [-2 2 -3 1/2] → bFcf
{74,240}{71,120}{65,120} → [-3 1/2 -6 1] → cffZ
ENCODING: CFBf BfBD BDAh AhaF aFbF bFcF cffZ

Fig. 1. Example of a 3-word encoding of a MIDI file with 120 ticks per beat resolution. Sequences of pairs {*MIDI pitch, duration in ticks*} are extracted using a window of length 3. Then, intervals and IOR within the window are calculated and finally are encoded using the encoding scheme.

Table 1. Number of words in the training sets for the different word lengths: number of different words in each style, total of different words found in each corpus, and coverage of the vocabulary (percentage).

	Corpus C1				Corpus C2				
n	Jazz	Clas.	$\lvert\mathcal{V}_n\rvert$	Coverage (%)	Bach	Greg.	Joplin	$\lvert\mathcal{V}_n\rvert$	Coverage (%)
2	425	485	548	49,24	245	87	223	301	27,04
3	4883	3840	7903	0,64	1374	509	1471	2605	0,21
4	6481	6209	12501	$9,07 \cdot 10^{-4}$	2574	1207	2708	5835	$4,23 \cdot 10^{-4}$
5	6849	7390	14198	$9,25 \cdot 10^{-7}$	3245	1901	3245	8073	$5,26 \cdot 10^{-7}$
6	6967	8060	15013	$8,79 \cdot 10^{-10}$	3594	2297	3420	9165	$5,37 \cdot 10^{-10}$
7	7018	8483	15499	$8,15 \cdot 10^{-13}$	3728	2413	3463	9564	$5,03 \cdot 10^{-13}$

2.3 Word Lengths

In order to test the classification ability of different word lengths, n, a range for $n \in \{2, 3, 4, 5, 6, 7\}$ has been established. The shorter n-words are less specific and provide more general information and, on the other hand, larger n-words may be more informative but the models based on them will be more difficult to train. Using the encoding scheme, the vocabulary size for each n is $|\mathcal{V}_n| = (53 \times 21)^{n-1}$ words. In Table 1 the number of words that have been extracted from the training set for each length is displayed.

2.4 Naive Bayes Classifier

The naive Bayes classifier, as described in [8], has been used. In this framework, classification is performed following the well-known *Bayes' classification rule*. In a context where we have a set of classes $c_j \in \mathcal{C} = \left\{c_1, c_2, \ldots, c_{|\mathcal{C}|}\right\}$, a melody x_i is assigned to the class c_j with maximum a posteriori probability, in order to minimize the probability of error:

$$P(c_j|x_i) = \frac{P(c_j)P(x_i|c_j)}{P(x_i)} \ . \tag{3}$$

Our classifier is based on the *naive Bayes assumption*, i.e. it assumes that all words in a melody are independent of each other, and also independent of the order they are generated. This assumption is clearly false in our problem and also in the case of text classification, but naive Bayes can obtain near optimal classification errors in spite of that [9]. To reflect this independence assumption, melodies can be represented as a vector $x_i = \left(x_{i1}, x_{i2}, \ldots, x_{i|\mathcal{V}|}\right)$, where each component $x_{it} \in \{0, 1\}$ represents whether the word w_t appears in the document or not, and $|\mathcal{V}|$ is the size of the vocabulary. Thus, the class-conditional probability of a document $P(x_i|c_j)$ is given by the probability distribution of words w_t in class c_j, which can be learned from a labelled training sample, $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, using a supervised learning method.

Multivariate Bernoulli Model (MB). Using this approach, each class follows a multivariate Bernoulli distribution:

$$P(x_i|c_j) = \prod_{t=1}^{|\mathcal{V}|} x_{it} P(w_t|c_j) + (1 - x_{it})(1 - P(w_t|c_j)) \tag{4}$$

where $P(w_t|c_j)$ are the class-conditional probabilities of each word in the vocabulary, and these are the parameters to be learned from the training sample.

Bayes-optimal estimates for probabilities $P(w_t|c_j)$, with a Laplacean prior to smooth probabilities, and prior probabilities for classes $P(c_j)$, are calculated as:

$$P(w_t|c_j) = \frac{1 + M_{tj}}{2 + M_j} \qquad P(c_j) = \frac{M_j}{|\mathcal{X}|} \tag{5}$$

where M_{tj} is the number of melodies in class c_j containing word w_t, and M_j is the total number of melodies in class c_j.

Multinomial Model (MN). This model takes into account word frequencies in each melody, rather than just the occurrence or non-occurence of words as in the MB model. In consequence, documents are represented by a vector, where each component x_{it} is the number of occurrences of word w_t in the melody. In this model, the probability that a melody has been generated by a class c_j is the multinomial distribution, assuming that the melody length in words, $|x_i|$, is class-independent [8]:

$$P(x_i|c_j) = P(|x_i|)|x_i|! \prod_{t=1}^{|\mathcal{V}|} \frac{P(w_t|c_j)^{x_{it}}}{x_{it}!} \tag{6}$$

Now, Bayes-optimal estimates for class-conditional word probabilities are:

$$P(w_t|c_j) = \frac{1 + N_{tj}}{|\mathcal{V}| + \sum_{k=1}^{|\mathcal{V}|} N_{kj}} \tag{7}$$

where N_{tj} is the sum of occurrences of word w_t in melodies in class c_j. Class prior probabilities are also calculated as for MB.

Multivariate Bernoulli Mixture Model (MMB). Both MB and MN have proven to achieve quite good results in text classification [8, 10], but they can be improved assuming that probabilities of words follow a more complex distribution within a class. Imagine that we want to model the distribution of *musical words* (see 2.2) in a sample of classical music melodies, which is formed in equal shares by Baroque and Renaissance scores. It is likely that both subsets have different distributions of words, so that we could find some words very common in Baroque music that don't usually appear in Renaissance music. In this case, using the MB model, Bayes-optimal estimates for these words in the whole set would be just a half of the real ones in Baroque, and much greater than their real value in Renaissance.

Intuitively, it would be more accurate to find the separate estimates for each substyle and then combine them to model the whole style. However, since the only information we have about these sequences is that all of them belong to classical music, finding the optimal estimates for each substyle is not straightforward. Furthermore, this problem becomes more complex when inner class structure (i.e. the number and proportions of the subsets) is not known a priori. To solve this problem, we will use here finite mixtures [11] of multivariate Bernoulli distributions, as they have been successfully applied in text classification tasks [10, 12].

A finite mixture model is a probability distribution formed by a number of components M:

$$P(x_i) = \sum_{m=1}^{M} \pi_m P(x_i|p_m) \tag{8}$$

where π_m are the mixing proportions, that must satisfy the restriction $\sum_{m=1}^{M} \pi_m = 1$; and $p_m = \left(p_{m1}, p_{m2}, \ldots, p_{m|\mathcal{V}|}\right)$ are the component prototypes. Since we will model each class as a mixture of multivariate Bernoulli distributions, each component distribution $P(x_i|p_m)$ is calculated using Eq. 4, substituting $P(w_t|c_j)$ with its corresponding value p_{mt}.

Now we face the problem of obtaining optimal estimates for parameters $\Theta = (\pi_1, \ldots, \pi_M, p_1, \ldots, p_M)^t$. This can be achieved using the EM algorithm [13], but to be applicable, it requires that the problem be formulated as an incomplete-data problem. To do this, we can think of each sample document x_i as an incomplete vector [12], where $z_i = (z_{i1}, \ldots, z_{iM})$ is the *missing data* and indicates which component of the mixture the document belongs to (with 1 in the position corresponding to the component and zeros elsewhere). Then, the EM proceeds iteratively to find the parameters that maximize the log-likelihood of the complete data:

$$\mathcal{L}_C(\Theta|X,Z) = \sum_{i=1}^{n}\sum_{m=1}^{M} z_{im}\left(\log \pi_m + \log P(x_i|p_m)\right) . \tag{9}$$

2.5 Feature Selection

The methods explained above use a representation of musical pieces as a vector of symbols. A common practice in text classification is to reduce the dimensionality of those vectors by selecting the words which contribute most to discriminate the class of a document. A widely used measure to rank the words is the *average mutual information* (AMI) [14].

For the MB model, the AMI is calculated between (1) the class of a document and (2) the absence or presence of a word in the document. We define C as a random variable over all classes, and F_t as a random variable over the absence or presence of word w_t in a melody, F_t taking on values in $f_t \in \{0,1\}$, where $f_t = 0$ indicates the absence of word w_t and $f_t = 1$ indicates the presence of word w_t. The AMI is calculated for each w_t as[1]:

$$I(C;F_t) = \sum_{j=1}^{|C|}\sum_{f_t\in\{0,1\}} P(c_j,f_t)\log\frac{P(c_j,f_t)}{P(c_j)P(f_t)} \tag{10}$$

where $P(c_j)$ is the number of melodies for class c_j divided by the total number of melodies; $P(f_t)$ is the number of melodies containing the word w_t divided by the total number of melodies; and $P(c_j,f_t)$ is the number of melodies in class c_j having a value f_t for word w_t divided by the total number of melodies.

In the MN model, the AMI is calculated between (1) the class of the melody from which a word occurrence is drawn and (2) a random variable over all the word occurrences, instead of melodies. In this case, Eq. 10 is also used, but $P(c_j)$ is the number of word occurrences appearing in melodies in class c_j divided by the total number of word occurrences, $P(f_t)$ is the number of occurrences of the word w_t divided by the total number of word occurrences, and $P(c_j,f_t)$ is the number of occurrences of word w_t in melodies with class label c_j, divided by the total number of word occurrences.

3 Results

For each model and word size, the naive Bayes classifier was applied to the n-words extracted from the melodies in our training set in order to test its style recognition ability. The experiments have been made following a leave-one-out scheme. The presented results are the percentage of successfully classified melodies.

The evolution of the classification as a function of the significance of the information used is presented in the graphs in Fig. 2. For this, the words in the training set have been ordered according to their AMI value. Also, experiments using only the best rated words ($|\mathcal{V}|$ in the graphs) have been performed.

[1] The convention $0 \log 0 = 0$ was used, since $x \log x \to 0$ as $x \to 0$.

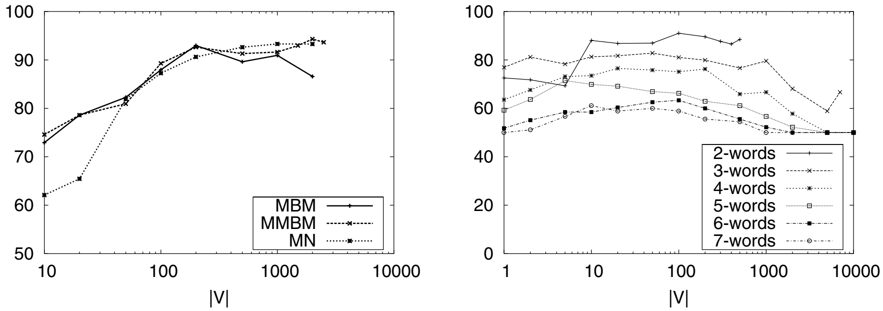


Fig. 2. Evolution of style recognition percentage in average for different vocabulary sizes. The plots represent **(left)** a comparison of the different statistical models (displayed for corpus C1) and **(right)** word sizes (for corpus C2).

Note that the results were not conclusive in terms of the different statistical models, since all the methods performed comparatively. There is a tendency of Bernoullis to classify better for small values of $|\mathcal{V}|$ while multinomials seem to provide better results for larger $|\mathcal{V}|$.

Table 2 shows the best results obtained in the experiments. The best accuracy was obtained for the word size $n = 3$, reaching a 94.3% of successful style identification. Large n-words only perform well for very small $|\mathcal{V}|$ values, and get worse rapidly for larger values. This preference for little specific information points to the fact that the training set is small for those lengths, and the results could be improved for larger models with more training melodies.

Also the values for precision and recall have been studied. Recall figures get very low as n increases, being the cause of the lower classification rates obtained for large words. In fact, the tendency for lengths $n = 2, 3$ is to get low percentage rates when $|\mathcal{V}|$ increases due to low recall and very high precision values: there are a lot of unclassified melodies, but the decisions taken by the classifier are usually very precise. It can be said that the classifier learns very well but little. This fact also reflects the need of a larger training set for these lengths to be successfully applied.

We have compared our results to those obtained by our research group with corpus C1, using melodic, harmonic and rhythmic statistical descriptors. They are fed into supervised classifiers like nearest neighbours (NN) or a standard Bayes rule [5]. In those experiments, the best recognition rates obtained for whole melodies were 91.0% for Bayes and 93.0% for NN, after a long study of the parameter space and the descriptor selection procedures. Thus, the first results obtained under this new approach are very encouraging.

4 Conclusions

The feasibility of using text categorization technologies for music style recognition has been tested. The models based on 2-words had the best performance,

Table 2. Best results in classification percentages obtained for both corpora. For each word length value, n, the table shows, from left to right: best classification, statistical model, size of vocabulary used for it and number of components (for the MMB).

	Corpus C1				Corpus C2			
n	% ok	model	$\|\mathcal{V}\|$	M	% ok	model	$\|\mathcal{V}\|$	M
2	92.1	MMB	100	2	90.6	MN	100	-
3	89.1	MMB	7500	4	94.3	MMB	2000	3
4	80.7	MB	100	-	90.6	MB	1000	-
5	82.3	MN	200	-	84.6	MN	7000	-
6	78.8	MN	14000	-	69.5	MN	50	-
7	73.9	MN	10000	-	53.3	MB	500	-

although the best result obtained was for $n = 3$, reaching a 94.3% of successful style recognition. Larger word lengths have provided also good results using small vocabulary sizes. In these cases, the method has proved to be very accurate, but lacks retrieval power. This fact points to a lack of training data. The various statistical models tested did not present significant differences in classification.

The results have been compared to those obtained by other description and classification techniques, providing similar or even better results. We are convinced that an increment of the data available for training will improve the results clearly, specially for larger n-word sizes.

Acknowledgment

This work was supported by the projects Spanish CICyT TIC2003–08496–C04, EU ERDF funds, and Generalitat Valenciana GV043–541.

References

1. Zhu, J., Xue, X., Lu, H.: Musical genre classification by instrumental features. In: Int. Computer Music Conference, ICMC 2004. (2004) 580–583
2. Whitman, B., Flake, G., Lawrence, S.: Artist detection in music with minnowmatch. In: Proc. of 2001 IEEE Workshop on Neural Networks for Signal Processing. (2001) 559–568
3. Soltau, H., Schultz, T., Westphal, M., Waibel, A.: Recognition of music types. In: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1998). (1998)
4. Cruz, P.P., Vidal, E., Pérez-Cortes, J.C.: Musical style identification using grammatical inference: The encoding problem. In Sanfeliu, A., Ruiz-Shulcloper, J., eds.: Proc. of CIARP 2003. (2003) 375–382
5. Ponce de León, P.J., Iñesta, J.M.: Feature-driven recognition of music styles. In: 1st Iberian Conference on Pattern Recognition and Image Analysis. LNCS, 2652. (2003) 773–781
6. McKay, C., Fujinaga, I.: Automatic genre classification using large high-level musical feature sets. In: Int. C. Music Information Retrieval, ISMIR'04. (2004) 525–530

7. Doraisamy, S., Rüger, S.: Robust polyphonic music retrieval with n-grams. Journal of Intelligent Information Systems **21** (2003) 53–70
8. McCallum, A., Nigam, K.: A comparison of event models for naive bayes text classification. In: AAAI-98 W. on Learning for Text Categorization. (1998) 41–48
9. Domingos, P., Pazzani, M.: Beyond independence: conditions for the optimality of simple bayesian classifier. Machine Learning **29** (1997) 103–130
10. Novovičová, J., Malík, A.: Text document classification using finite mixtures. Technical Report 2063, Academy of Sciences of the Czech Republic, Institute of Information Theory and Automation (2002)
11. McLachlan, G., Peel, D.: Finite Mixture Models. John Wiley & Sons (2000)
12. Juan, A., Vidal, E.: On the use of Bernoulli mixture models for text classification. Pattern Recognition **35** (2002) 2705–2710
13. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. of the Royal Statistical Society B **39** (1977) 1–38
14. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley (1991)

The MORFO3D Foot Database*

José García-Hernández[1], Stella Heras[1], Alfons Juan[1],
Roberto Paredes[1], Beatriz Nácher[2], Sandra Alemany[2],
Enrique Alcántara[2], and Juan Carlos González[2]

[1] Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
{jogarcia,sheras,ajuan,rparedes}@dsic.upv.es
[2] Institut de Biomecànica de València
Universitat Politècnica de València
{beanacfe,maalmut,ealcanta,jcgonza}@ibv.upv.es

Abstract. A foot database comprising 3D foot shapes and footwear fitting reports of more than 300 participants is presented. It was primarily acquired to study footwear fitting, though it can also be used to analyse anatomical features of the foot. In fact, we present a technique for automatic detection of several foot anatomical landmarks, together with some empirical results.

1 Introduction

Footwear fitting has a decisive influence on the functionality and comfort of shoes. The conventional technique to estimate footwear fitting is the direct try of shoes, which is perfectly possible in most cases. However, there are many cases in which it is not possible due to the amount of models to try, physical obstacles to the technique (e.g. in e-commerce), or physical or psychical problems related with the user (e.g. sensitivity difficulties). In all of these cases, *automatic fitting prediction* can help a lot.

In this work, we present a foot database developed in the Spanish research project MORFO3D. The MORFO3D foot database was acquired during the month of May 2004 at the *Universitat Politècnica de València,* València (Spain). It comprises 3D foot shapes and footwear fitting reports of more than 300 participants. The database was primarily acquired to study footwear fitting, though it can also be used to analyse anatomical features of the foot. In fact, we present a knowledge-based technique for automatic detection of several foot anatomical landmarks that gives good results. The MORFO3D foot database is available upon request for non-commercial use.

The acquisition of the MORFO3D foot database is described in section 2. Then, in section 3, we present our technique for automatic detection of several foot anatomical landmarks, together with some empirical results.

* Work supported by the "Agència Valenciana de Ciència i Tecnologia" under grant GRUPOS03/031 and the Spanish projects DPI2001-0880-CO2-01 and DPI2001-0880-CO2-02.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 658–665, 2005.

2 The MORFO3D Foot Database

The MORFO3D foot database was acquired during the month of May 2004 at the *Universitat Politècnica de València,* València (Spain). A total of 316 18- to 35-year-old women of European shoe size 38 (and 40) participated in its acquisition. It is considered that the feet of the women in this age interval is enough developed, but still not shows the typical advanced age pathologies. For each participant, we first acquired her right 3D foot shape together with the location of several foot key points (landmarks). Then, we asked her to try 4 models of shoes (out of 8 available) and fill in a questionnaire about each shoe fitting. Hereafter we provide a detailed description of these two basic acquisition steps.

For the acquisition of 3D foot shapes, we used an INFOOT laser scanner [1]. This scanner is able to acquire a complete 3D shape of the foot, including the foot sole. Also, it can be used to acquire the location of foot landmarks by simply marking them on the user barefoot with adhesive markers. In our case, the acquisition process is as illustrated in Figure 1. The process begins with the placement of adhesive markers on 14 foot landmarks located on bony prominences or critical zones for shoe fitting (see top of Figure 1 for their precise locations). Then, the scanning process is carried out while the participant stands upright with equal weight on each foot, in a certain position and orientation (see middle of Figure 1). The result consists of: a) a cloud of points representing the outer surface of the foot; b) the location of the 14 foot landmarks previously marked; and c) a number of podometric measurements derived from these 14 landmarks (see bottom of Figure 1). The complete acquisition process lasted approximately 5 minutes on average.

Table 1 shows some descriptive statistics of the 3D foot shapes acquired. The main differences among the feet measurements taken were observed in the height of the external malleolus and the foot length.

Table 1. Descriptive statistics (in mm) of the 3D foot shapes acquired

Podometric variable	Min	Mean	Max	Standard deviation
Foot length	225.9	241.77	257.6	±5.82
Forefoot width	84.5	94.07	106.4	±3.95
Hell width	55.2	61.44	70.5	±2.88
Instep height	51.8	62.56	80.8	±4.29
1st toe height	7.6	16.92	27.7	±3.43
Height of the external malleolus	49.9	64.52	122.7	±7.93

The second basic acquisition step was designed to compile detailed information about the fitting of some shoes to the participant's feet. We purchased several pairs of 8 different shoe models that we thought representative of the models available in the market during the database acquisition (see top of Figure 2). Given a pair of shoes (and socks), the participant was asked to try it

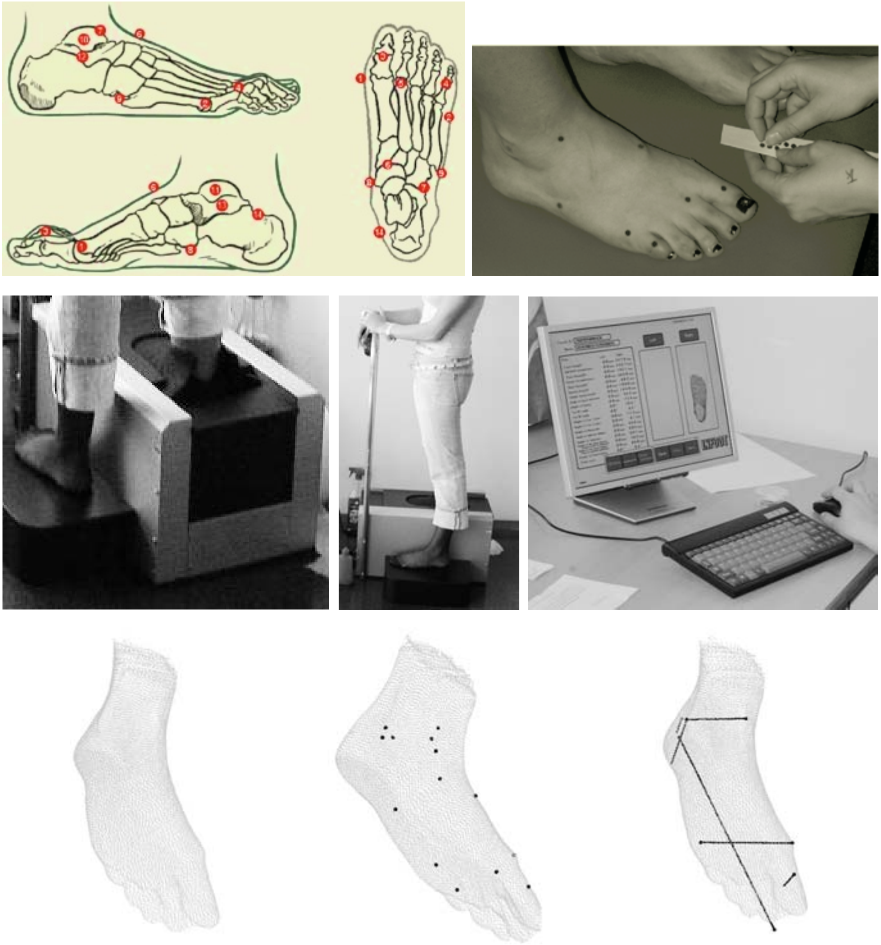


Fig. 1. Acquisition of 3D foot shapes. Top-left: locations of the 14 foot landmarks; 1) Metatarsal Tibiale, 2) Metatarsal Fibulare, 3) Highest point of the 1st toe at the interphalangeal joint, 4) Highest point of the 5th toe joint at the distal interphalangeal joint, 5) Head of the second metatarsal, 6) Instep point (Cuneiform), 7) Tentative junction point, 8) Navicular, 9) Tuberosity of five metatarsal, 10) The most lateral point of lateral malleolus, 11) The most medial point of medial malleolus, 12) Sphyrion Fibulare, 13) Sphyrion, and 14) Medial tentative heel upper point. Top-right: placement of adhesive markers at the locations of the 14 foot landmarks. Middle: scanner parts and participant's position and orientation during the scanning process. Bottom: output of the process; from left to right: 3D foot shape described as a cloud of points, locations of the 14 landmarks, and 6 podometric measurements derived from these landmarks (foot length, forefoot width, hell width, instep height, 1st toe height and height of the external malleolus)

walking during 2 minutes and then fill in a questionnaire about the fitting in 15 different zones of the foot (see bottom of Figure 2). For each of these zones, the participant gave her perception of discomfort in a 4-level ordinal scale (0=none, 1=low, 2=medium, 3=high). Also, the participant answered questions about the global discomfort of the shoes and her general footwear preferences. This step lasted 5 minutes on average for the trial of each model. Unfortunately, the high time cost of the complete trial (40 minutes for 8 models) prevented trying all the available shoe models, so we decided to provide the participant with only 4 randomly selected models. Therefore, on average, each shoe model was tried by half of the 316 participants.

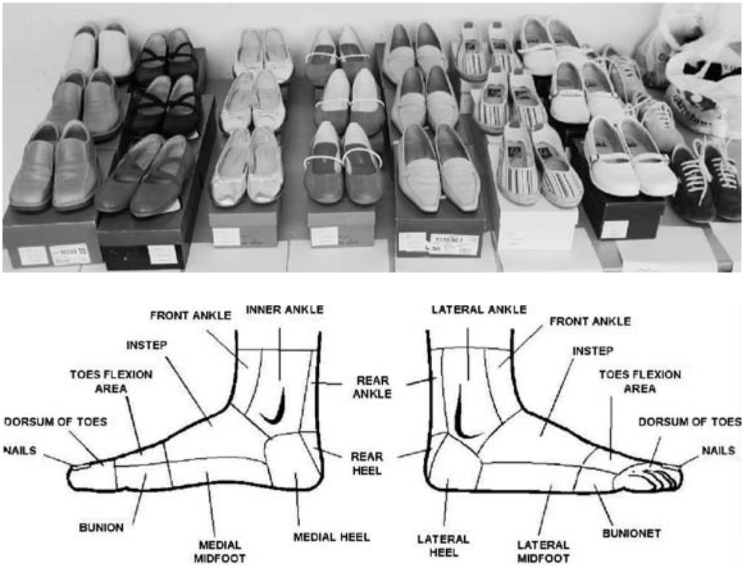


Fig. 2. Top: shoe models purchased. Bottom: zone division of the foot

Table 2 shows the mean and the standard deviation of the discomfort variable for each shoe model and foot zone.

The MORFO3D foot database is available upon request for non-commercial use. The complete database also stores data of 100 women right foot of European shoe size 40. The data in the MORFO3D database is being used to study the relation between footwear fitting and podometric measurements.

3 Automatic Detection of Landmarks

The *classical method* to detect foot anatomical landmarks described in section 2 requires an expert to manually place them. The landmark placement obtained from different experts, or from the same expert at different moments, can be

Table 2. Mean and standard deviation of the discomfort variable (0=none, 1=low, 2=medium, 3=high) for each shoe model and foot zone

		Shoe model								
	Foot zone	1	2	3	4	5	6	7	8	Average
1	Rear heel	0.3±0.7	0.9±1.1	0.3±0.7	0.2±0.5	0.5±0.8	0.3±0.7	0.5±0.8	0.5±0.8	0.5±0.8
2	Lateral heel	0.2±0.5	0.7±1.0	0.2±0.6	0.1±0.4	0.3±0.7	0.2±0.6	0.3±0.6	0.3±0.7	0.3±0.7
3	Medial heel	0.2±0.5	0.7±1.0	0.2±0.5	0.1±0.3	0.3±0.6	0.2±0.5	0.3±0.6	0.3±0.6	0.3±0.7
4	Rear ankle	0.3±0.6	0.5±1.0	0.2±0.6	0.2±0.6	0.3±0.7	0.2±0.6	0.4±0.8	0.6±0.9	0.3±0.7
5	Lateral ankle	0.2±0.5	0.4±0.8	0.1±0.4	0.1±0.2	0.1±0.5	0.1±0.4	0.2±0.5	0.3±0.7	0.2±0.6
6	Inner ankle	0.2±0.5	0.3±0.8	0.1±0.4	0.1±0.2	0.1±0.3	0.1±0.4	0.3±0.6	0.4±0.7	0.2±0.5
7	Front ankle	0.2±0.5	0.3±0.7	0.1±0.4	0.1±0.4	0.1±0.4	0.1±0.4	0.1±0.4	0.4±0.8	0.2±0.5
8	Instep	0.6±0.8	1.4±1.1	0.1±0.5	0.5±0.8	0.3±0.6	0.2±0.5	0.9±1.1	0.8±0.9	0.6±0.9
9	Lateral midfoot	0.3±0.6	0.9±1.0	0.2±0.5	0.2±0.5	0.2±0.6	0.1±0.4	0.4±0.7	0.5±0.8	0.3±0.7
10	Medial midfoot	0.2±0.6	0.8±1.0	0.2±0.5	0.2±0.5	0.2±0.5	0.1±0.4	0.3±0.7	0.4±0.8	0.3±0.7
11	Toes flexion area	0.3±0.6	1.3±1.1	0.3±0.7	0.2±0.5	0.5±0.7	0.2±0.5	1.2±1.1	0.5±0.8	0.6±0.9
12	Bunionette	0.2±0.5	1.0±1.1	0.3±0.6	0.1±0.4	0.3±0.7	0.2±0.4	0.5±0.8	0.4±0.7	0.4±0.7
13	Bunion	0.2±0.5	0.9±1.1	0.3±0.7	0.1±0.4	0.3±0.7	0.2±0.5	0.5±0.8	0.3±0.6	0.4±0.7
14	Dorsum of toes	0.4±0.7	1.4±1.2	1.1±1.1	0.2±0.5	1.1±1.1	0.5±0.8	0.8±1.1	0.5±0.8	0.7±1.0
15	Nails	0.4±0.8	1.2±1.2	1.3±1.1	0.2±0.5	0.9±1.1	0.7±0.9	0.9±1.2	0.5±0.8	0.8±1.0
	Average	0.3±0.4	0.9±0.7	0.3±0.4	0.2±0.3	0.4±0.4	0.2±0.3	0.5±0.4	0.5±0.4	0.4±0.5

different. In this work, we propose a technique to automatically perform this landmark placement.

The *technique* proposed is based on heuristics and takes into account the fact that the scanner places automatically the reference axis shown in the top-right of the Figure 3. It works initially with the complete 3D point clouds obtained from the scanner and can be described as follows. First, a set of initial landmarks is constructed from the complete cloud. It includes the landmarks shown as points 6, 16, 17 and 21 in the top-left and top-center of the Figure 3. For instance, the landmark 17 is the point with a smaller x-coordinate from the complete 3D cloud and the landmark 21 is the point with y-coordinate$\simeq 0$, z-coordinate$\simeq 0$ and the greatest x-coordinate. The set of initial landmarks allows us to split the cloud along the x and z-axes into three sub-clouds: upper, rear-lower and fore-lower subclouds. Then, more specific points are searched in each sub-cloud. For instance, in rear-lower subcloud, the points 19 and 20 from Figure 3 are located as those points with, respectively, the smallest and the greatest y coordinate. Working in this way, the points 1, 2, 15 and 18 are detected in the fore-lower subcloud, and the points $10 - 13$ in the upper subcloud. Note that most of the automatic landmarks are border points, that is, they are coordinate maximums or minimums in an axis of a sub-cloud. Only the landmarks 6, 12, 13 and 16 are not border points. The landmark 6 (instep point) is the intersection point of the cloud with the line that goes through the midpoint between landmarks 17 and 21, in the direction of the Z axis. The landmarks 12, 13 and 16 are heuristically detected from, respectively, the landmarks 10, 11 and 17.

The set of landmarks detected automatically does not match exactly with that obtained manually (see Figure 1). Some landmarks that are located manually can be automatically detected, but not all of them. For instance, the landmark points 8 and 9 of the top-left and top-center of Figure 3 show the joint point of two bones. Therefore, they can only be placed by touching the foot, since they can not be seen as any protuberance.

We have compared the landmark set obtained manually with the landmark set obtained using our automatic technique. As both sets do not completely match, we have compared their intersection. That is, given one foot, each auto-

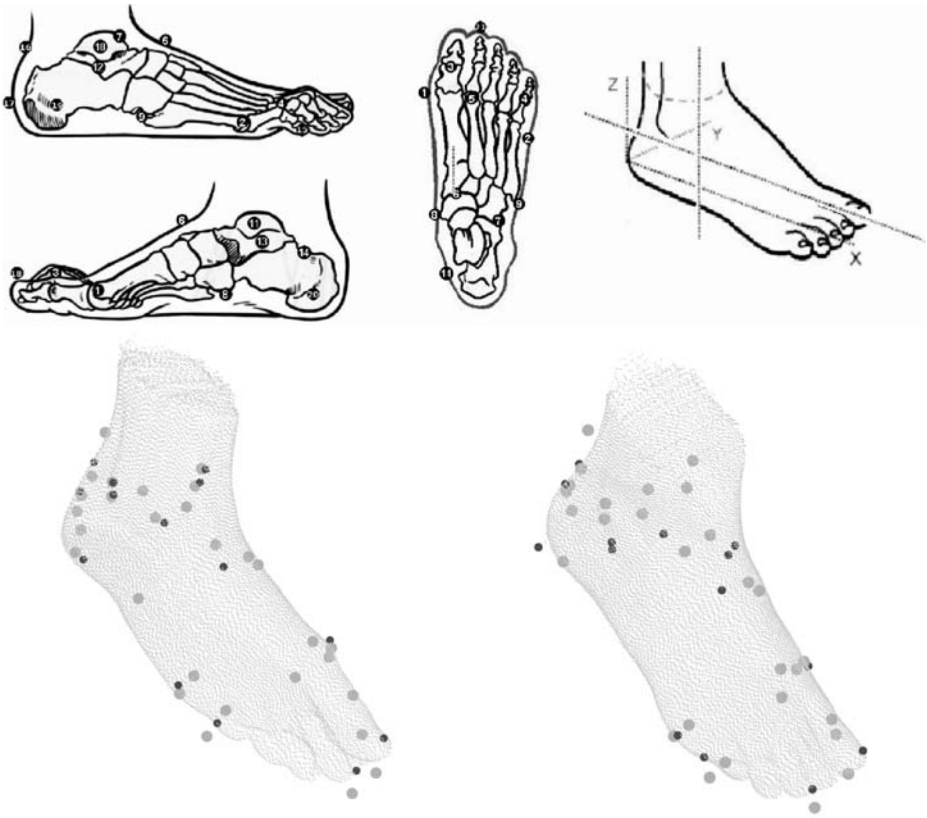


Fig. 3. Automatic detection of landmarks. Top-left and top-center: landmarks detected by the automatic method; 1) Metatarsal Tibiale, 2) Metatarsal Fibulare, 6) Instep point (Cuneiform), 10) The most lateral point of lateral malleolus, 11) The most medial point of medial malleolus, 12) Sphyrion Fibulare, 13) Sphyrion, 15) Most lateral point of the 5th toe, 16) Insertion of Achille's tendom in calcaneus, 17) Heel rearrest point, 18) Highest point of the 1st toe, 19) Most prominent point of the external heel, 20) Most prominent point of the internal heel and 21) Most advanced point of the 2nd toe. Landmarks numbered from 1 to 14 are those detected by the classical method (see Figure 1 for more details). Top-right: reference axis with lines that split the point clouds in three subclouds. Bottom : the automatic method application. Black points are manually placed by the expert, and grey points are detected automatically. In the left image the automatic method works properly, but in the right one the method fails

matic landmark from intersection points set has been compared with its respective manual landmark. We have use different approaches for it. On one hand, we have use three distance definitions: L_1, L_2 (or Euclidean Distance) and mean squared error (MSE). The L_1 and L_2 distances from a manual landmark x to its corresponding automatic landmark y for a given foot are shown in the equations 1 and 2. The results shown in Table 3 are the distance averages. The

MSE from the set of observed values (manual landmarks) $X = (x_1, x_2, \ldots, x_N)$ to calculated values (automatic landmarks) $Y = (y_1, y_2, \ldots, y_N)$ is given in the equation 3. The result is shown in Table 3.

$$L_1(x, y) = |x_1 - y_1| + |x_2 - y_2| + |x_3 - y_3| \tag{1}$$

$$L_2(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2} \tag{2}$$

$$MSE(X, Y) = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}} \tag{3}$$

On the other hand, Table 4 shows the error rate for different thresholds of tolerance. It is computed as follows: for each threshold value, we say that an automatic landmark matches with its corresponding manual landmark when the L_2 distance between them is smaller than the given threshold value (success); otherwise, we say that they do not match (error).

The bottom of Figure 3 shows two examples of application of our automatic technique. In the examples, the automatic landmarks are compared with the classical landmarks. In the left image, our technique works properly but, in the right one, it did not place the anatomical points in the correct position.

Besides, the automatic landmark set of a foot can be used to compute a minimal but complete set of automatic podometry features, those shown in Table 1. These features can be used to automatically characterize the foot.

Table 3. Different error rates (in mm) for each point of the intersection set and total error rate. Numbers in brackets correspond with those of the Figure 3

Point	L_1	L_2	MSE
Lateral malleolus (10)	33.68	22.39	5.8
External heel (19)	60.63	46.18	7.78
Medial malleolus (11)	43.77	30.02	6.62
Internal heel (20)	41.89	29.80	6.47
Heel rear (17)	18.15	14.01	4.26
Total	39.62	28.48	6.19

Table 4. Error rate (%) for different tolerance thresholds (distance in mm between classical and automatic anatomical points)

Distance	0	1	2	3	4	5	6
Error	98	90	76	66	60	56	52
Distance	7	8	9	10	11	12	13
Error	49	47	44	42	40	38	36
Distance	14	15	16	17	18	19	20
Error	34	33	32	31	30	29	28

4 Conclusions

A foot database comprising 3D foot shapes and footwear fitting reports of more than 300 participants has been presented. It is called MORFO3D foot database. It can be used to study footwear fitting, and also to analyse anatomical features of the foot. In fact, we have presented a technique for automatic detection of foot anatomical landmarks that gives good results. The MORFO3D foot database is available upon request for non-commercial use.

References

1. I-ware laboratory. http://www.i-ware.co.jp/.
2. R. S. Goonetilleke and A. Luximon. Designing for comfort: a footwear application. In *Proceedings of the Computer-Aided Ergonomics and Safety Conference'01*, July 28-Aug 2 2001. Plenary session paper.
3. B. Nacher, E. Alcántara, S. Alemany, J. García-Hernández, and A. Juan. 3d foot digitalizing and its application to footwear fitting. In *In Proc. of 3D Modelling*, 2004.

Fast Surface Grading Using Color Statistics in the CIE Lab Space*

Fernando López[1], José Miguel Valiente[1], Ramón Baldrich[2], and María Vanrell[2]

[1] Universidad Politécnica de Valencia, 46022 Valencia, Spain
flopez@disca.upv.es
[2] Universitat Autònoma de Barcelona, CVC, 08193 Cerdanyola del Vallès, Spain

Abstract. In this paper we approach the problem of fast surface grading of flat pieces decorated with random patterns. The proposed method is based on the use of global statistics of color computed in the CIE Lab space. Two other fast methods based on color histograms [1] and Centile-LBP features [8] are introduced for comparison purposes. We used CIE Lab in order to provide accuracy and perceptual approach in color difference computation. Experiments with RGB were also carried out to study CIE Lab reliability. The ground truth was provided through an image database of ceramic tiles. Nevertheless, the approach is suitable to be extended to other random decorated surfaces like marble, granite, wood or textile stuff. The experiments make us to conclude that a simple collection of global statistics of color in the CIE Lab space is powerful enough to well discriminate surface grades. The average success surpasses 95% in most of the tests, improving literature methods and achieving factory compliance.

1 Introduction

The background problem is to solve the question of surface grading of flat pieces decorated with random patterns. These include surfaces from nature (wood, marble or granite) and artificial surfaces (ceramic tiles or textile stuff). The aim of surface grading is to split the production into different classes sorted by their global appearance, which is crucial to achieve competitive quality standards. Industries related with the manufacturing of these products rely the task of grading on human operators. This grading is subjective and often inconsistent between different graders [7]. Thus, automatic and reliable systems are needed. Also, real time compliance is important in order to make systems able to inspect the overall production at on-line rates.

In the last decade many approaches about surface grading were developed, mainly for the industrial sectors of ceramics, marble, granite and wood. Boukouvalas et al [1][2][3] proposed color histograms and dissimilarity measures of these distributions to grade ceramic tiles. No real time compliance was studied.

Other works were related with an specific type of ceramic tiles, the *polished porcelanic* tiles, which imitate granite appearance. These works included texture

* Partially supported by FEDER-CICYT(DPI2003-09173-C02-01) and Keraben S.A.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 666–673, 2005.

features. Baldrich et al [4] proposed a perceptual approximation based on the use of discriminant features defined by human classifiers at factory. These features were mainly related to grain distribution and size. The method included grain segmentation and features measurement. Lumbreras et al [5] joined color and texture through multiresolution decompositions on several color spaces. They tested combinations of multiresolution decomposition schemes (Mallat's, *àtrous* and wavelet packets), decomposition levels and color spaces (Grey, RGB, Ohta and Karhunen-Loève transform). Peñaranda et al [6] used the first and second histogram moments of each channel of the RGB space. This simple approximation, together with a deep studied inspection system, were able to comply time requirements for on-line inspection. In Baldrich and Lumbreras's works there are no study about time compliance.

On wood grading, Kauppinnen [7] developed a method based on the percentile features of histograms calculated for RGB channels. These features are also called Centiles. Kyllönen et al [8] made an approach using color and texture features. For color they chose the above mentioned Centiles, and LBP (Local Binary Pattern) histograms for texture description.

Lebrun and Macaire [9] described the surfaces of the Portuguese "Rosa Aurora" marble using the mean color of the background and mean color, absolute density and contrast of marble veins. They achieved good results but their approach is very dependent on the properties of this marble. Finally, Kukkonen et al [10] presented a system for the grading of ceramic tiles using spectral images. Spectral images have the inconvenient of producing great amounts of data.

Table 1. Summary of surface grading literature.

	ground truth	**features**	**time study**	**accuracy %**
Boukouvalas	ceramic tiles	color	no	-
Baldrich	polished tiles	color/texture	no	92.0
Lumbreras	polished tiles	color/texture	no	93.3
Peñaranda	polished tiles	color/texture	yes	-
Kauppinen	wood	color	yes	80.0
Kyllönen	wood	color/texture	no	-
Lebrun	marble	color/texture	no	98.0
Kukkonen	ceramic tiles	color	no	80.0

Many of these approaches were very specialized in a specific type of surface, others did not achieve good enough accuracy, and others did not take into account the time restrictions of a real inspection at factory. As a result of this, we think surface grading is still an open research field. In this paper we present a generic method suitable to be used in a wide range of random surfaces; ceramic tiles, marble, granite, wood, textile stuff, etc. The approach uses fast and simple statistics of color, achieving good results with a representative data set of ceramic tiles. Thus, the method is appropriate to be implemented on systems with real time requirements, typical in these contexts.

2 Lab Statistics

The presented method is simple, a set of statistical features describing color properties are collected. The features are computed in a perceptually uniform color space, the CIE Lab. These statistics form a feature vector used in the classification stage where the well known k-NN method [11] was chosen as classifier.

CIE Lab was designed to be perceptually uniform. The term 'perceptual' is refered to the way that humans perceive colors, and 'uniform' implies that the perceptual difference between two coordinates (two colors) will be related to a measure of distance, which commonly is the Euclidean distance. Thus, color differences can be measured in a way close to the human perception of colors.

The images of the data set were acquired originally in RGB, then, conversion to CIE Lab coordinates was needed. This conversion is made using the standard RGB to CIE Lab transformation [12] as follows.

RGB to XYZ:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.412453 \; 0.357580 \; 0.180423 \\ 0.212671 \; 0.715160 \; 0.072169 \\ 0.019334 \; 0.119193 \; 0.950227 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

XYZ to CIE Lab:

$$\begin{aligned} L &= 116(Y/Y_n)^{1/3} - 16 \\ a &= 500((X/X_n)^{1/3} - (Y/Y_n)^{1/3}) \\ b &= 200((Y/Y_n)^{1/3} - (Z/Z_n)^{1/3}) \end{aligned}$$

X_n, Y_n, and Z_n are the values of X, Y and Z for the illuminant (reference white point). We followed the ITU-R Recommendation BT.709, and used the illuminant D_{65}, where $[X_n \, Y_n \, Z_n] = [0.95045 \;\; 1 \;\; 1.088754]$.

We proposed several statistical features for describing surface appearance. For each channel we chose the mean, the standard deviation $\sigma(z)$ and the average deviation $ADev(z)$.

$$\sigma(z) = \sqrt{\frac{\sum_{i=1}^{L}(z_i - m)}{L-1}} \qquad ADev(z) = \tfrac{1}{L}\sum_{i=1}^{L}|z_i - m|$$

where z is the random variable, L size of the data set and m the mean value of z values.

Also, by computing the histogram of each channel, we are able to calculate histogram moments. We defined two blocks of histogram moments; one from 2nd to 5th and the other from 6th to 10th. The nth moment of z about the mean is defined as

$$\mu_n(z) = \sum_{i=1}^{L}(z_i - m)^n p(z_i)$$

where z is the random variable, $p(z_i)$, $i = 1, 2, \ldots, L$ the histogram, L the number of distinct variable values and m the mean value of z.

3 Literature Methods

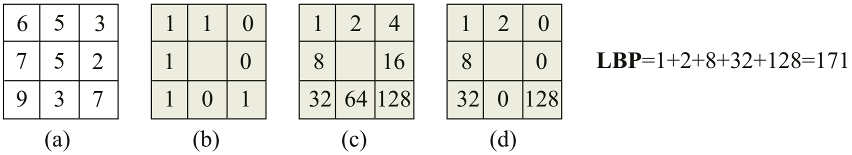
For comparison purposes we selected two methods from literature: color histograms [1] and Centile-LBP [8]. They are similar to ours, both are generic solutions with low computational costs. Color histograms are 3D histograms (one axis per space channel) which are compared using dissimilarity measures. In [1] they used the *chi square test* and the *linear correlation coefficient.*

$$\chi^2 = \sum_i \frac{(R_i - S_i)^2}{R_i + S_i} \qquad\qquad r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})}\sqrt{\sum_i (y_i - \bar{y})}}$$

When comparing two binned data sets with the same number of data points the *chi square* statistic (χ^2) is defined as above, where R_i is the number of events in bin i for the first data set, and S_i is the number of events in the same bin for the second data set. The *linear correlation coefficient* (r) measures the association between random variables for pairs of quantities *(x_i,y_i), i = 1,...,N.* The mean of the x_i values is $\bar{x}$ and $\bar{y}$ is the mean of the y_i values.

The Centiles, are calculated from a cumulative histogram $C_k(x)$, which is defined as a sum of all the values that are smaller than x or equal to x in the normalized histogram $P_k(x)$, corresponding to the color channel k. Finding a value for a percentile is finding the x when $C_k(x)$ is known, thus, requiring an inverse function of $C_k(x)$. Let $F_k(y)$ be the percentile feature, then $F_k(y) = C_k^{-1}(y) = x$, where y is a value of the cumulative histogram in the range [0%,100%].

The Local Binary Pattern (LBP) is a texture operator where the original 3x3 neighborhood is thresholded by the value of the center pixel (figure 1b). The values of the pixels in the thresholded neighbourhood are multiplied by the weights given to the corresponding pixels (figure 1c). Finally, the values of the eight pixels are summed to obtain the number of this texture unit. Using LBP there are 2^8 possible combinations of texture numbers, then a histogram collects the LBP texture description of an image.



6	5	3
7	5	2
9	3	7

(a)

1	1	0
1		0
1	0	1

(b)

1	2	4
8		16
32	64	128

(c)

1	2	0
8		0
32	0	128

(d)

LBP=1+2+8+32+128=171

Fig. 1. Computation of local binary pattern (LBP).

In [8] Centile and LBP features were combined in one measure of distance and then used the k-NN classifier. For Centile features they used the Euclidean distance in the feature space. For LBP they used a log-likelihood measure: $L(S,R) = -\sum_{n=0}^{N-1} S_n ln R_n$, where N is the number of bins. S_n and R_n are the sample and reference probabilities of bin n. The distances were joined by simply adding them. Previously both distances were normalized using the min and max values of all the distances found in the training set.

4 Experiments and Results

All the experiments were carried out using the same data set. The ground truth was formed by the digital RGB images of 492 tiles acquired from eight different models, each one with three different surface classes given by specialized graders at factory. For each model there were two close classes and one class far to them.

Models were chosen representing the extensive variety that factories can produce, a catalogue of 700 models is common. But, in spite of this great number of models, all of them imitate one of the following mineral textures; marble, granite or stone. Fixed pattern models are a subset of random pattern models.

Table 2. Ground truth of ceramic tiles.

	classes	tiles/class	size (cm)	pattern	aspect
Agata	13, 37, 38	16	33x33	fixed	marble
Berlin	2, 3, 11	24	16x16	random	granite
Firenze	9, 14, 16	20	20x25	random	stone
Lima	1, 4, 17	24	16x16	random	granite
Oslo	2, 3, 7	24	16x16	random	granite
Toscana	13, 18, 19	16	33x33	random	stone
Vega	30, 31, 37	20	20x25	fixed	marble
Venice	12, 17, 18	20	20x25	random	marble

Digital images of tiles were acquired using an illumination system spatially and temporally uniform. Spatial and temporal uniformity is important in surface grading [1,4,6] because variations on illumination can produce different shades for the same surface and then misclassifications. The illumination system was formed by two special high frequency fluorescent lamps with uniform illuminance along its length. For overcoming variations along time, the power supply is automatically regulated by a photoresistor located near fluorescents.

Two sets of experiments were made to demonstrate the feasibility of Lab statistics for solving the problem of surface grading. Firstly, experiments of statistics where carried out for the CIE Lab and RGB spaces. Classification was made using the half of the samples as training set and the remaining half as test set. Values of 1, 3, 5 and 7 were used for the k factor of the k-NN classifier.

The performance results of several statistics sets are shown in table 3. The error rates were computed as the average error ratios achieved over all models. More combinations of statistics were tested, but only the most prominent are presented. The last two columns corresponds to the averaged error rate and the 95% confidence intervals [11] respectively. The table is divided in two blocks, the first one corresponds with CIE Lab experiments. Here, the majority of sets have confidence intervals under the maximum error rate of 5% which is the factory requirement of performance. The best choice was to use the mean color plus the standard deviation. Histogram moments did not introduce any improvement. The second block collects the results of RGB which presents significant less discriminative power than CIE Lab.

Table 3. Accuracy results of statistics sets in CIE Lab and RGB spaces.

mean	stddev	avedev	2-5th ms	6-10th ms	lab	rgb	error %	95% c.i
x					x		13.2	[10.3, 16.4]
x	x				x		**1.2**	**[0.33, 2.3]**
x		x			x		**3.0**	**[1.6, 4.7]**
x	x		x		x		**3.2**	**[1.7, 4.9]**
x	x		x	x	x		3.3	[1.9, 5.2]
x						x	13.4	[10.4, 16.6]
x	x					x	7.9	[5.7, 10.6]
x		x				x	7.3	[5.1, 9.9]
x	x		x			x	5.9	[4.0, 8.3]
x	x		x	x		x	6.7	[4.6, 9.2]

Table 4. Accuracy results of Color Histograms and Centile-LBP.

	Chi	Corr	Log	Lab	RGB	error %	95% c.i
Color Histograms	x			x		9.7	[7.2, 12.6]
Color Histograms		x		x		11.5	[8.8, 14.6]
Color Histograms	x				x	11.1	[8.5, 14.2]
Color Histograms		x			x	12.4	[9.5, 15.5]
Centile-LBP	x			x		5.6	[3.6, 7.8]
Centile-LBP		x		x		5.1	[3.3, 7.4]
Centile-LBP			x	x		8.7	[6.4, 11.5]
Centile-LBP	x				x	5.3	[3.5, 7.6]
Centile-LBP		x			x	4.6	[2.8, 6.6]
Centile-LBP			x		x	6.7	[4.6, 9.2]

In second place, experiments for color histograms and Centile-LBP were carried out. Once again, classification was made using the half of the samples for training and the remaining half for testing. In Centile-LBP experiments the original log-likelihood formula, the *chi square test* and the *linear correlation coefficient* were used for measuring histograms differences.

The results of table 4 show that Centile-LBP achieves the best error rates when using RGB, but none of both methods achieves factory compliance because all of their confidence intervals surpass the max error rate of 5% required at factory. Comparing with table 3, Lab Statistics presents significant improvement in performance an also is the only method with confidence intervals complying the max factory error.

Figure 2 shows the best performance response of each method itemized by models. Color histograms and Centile-LBP approaches, contrasted with Lab Statistics, present greater irregularity and more models are over the factory max error.

Finally, we measured the timing costs of the methods using a common PC (see figure 3). All the approaches have a theoretical computational cost of $\Theta(n) + C$, where n is the image size and C is a constant which varies depending on the approach. Lab statistics and color histograms were penalized because they need

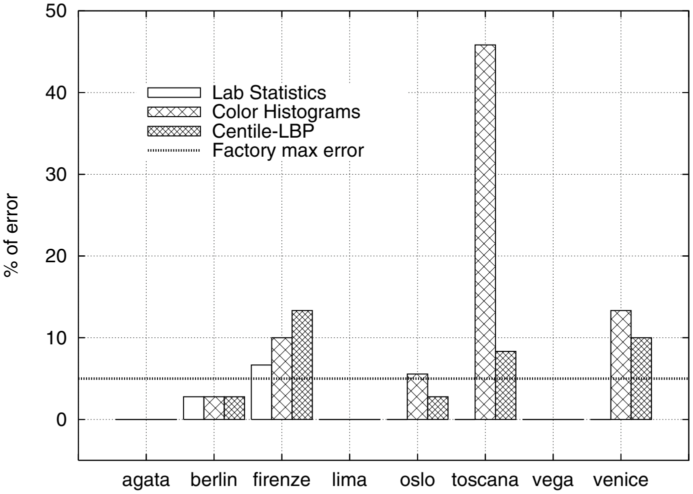


Fig. 2. Best accuracy results of Lab Statistics, Color Histograms, and Centile-LBP.

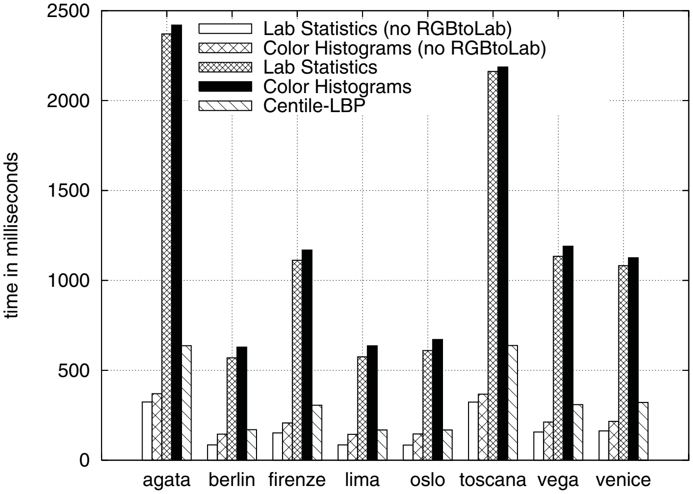


Fig. 3. Timing for the best accuracy results of each method.

the conversion from RGB to CIE Lab. But, if we take away the RGB to CIE Lab conversion then Lab statistics achieves the best timing response.

5 Conclusions and Further Work

A fast method for the application of surface grading has been presented. The method uses simple statistics computed in a perceptually uniform color space, the CIE Lab. This approach performs well discriminating correctly surface grades among several types of surfaces representing a common catalogue of ceramic tiles. The benefit of using CIE Lab is demonstrated comparing with RGB results.

Other two methods coming from the literature were implemented and tested for comparison purposes. From the point of view of performance, color histograms achieved the worse results while Centile-LBP had intermediate results. The best accuracy response corresponded to Lab statistics, which also was the only method achieving factory compliance in performance with confidence intervals under the max error limit. From the point of view of timing costs, Centile-LBP had the best response, but Lab statistics was not too far away and timing can be easily improved transferring the RGB to CIE Lab conversion to hardware or using parallel processing systems.

Further work will extend the image database with more models and samples. Also, a deep study of real time compliance will be made simulating factory load and using parallel processing systems based on cluster and MPI technology.

References

1. C. Boukouvalas et al. Color grading of randomly textured ceramic tiles using color histograms. IEEE Trans. Industrial Electronics. 46(1):219–226, 1999.
2. C. Boukouvalas and M. Petrou. Perceptual correction for colour grading using sensor transformations and metameric Data. Machine Vision and Applications, 11:96-104, 1998.
3. C. Boukouvalas and M. Petrou. Perceptual correction for colour grading of random textures. Machine Vision and Appl., 12:129-136, 2000.
4. R. Baldrich, M. Vanrell and J. J. Villanueva. Texture-color features for tile classification. EUROPTO/SPIE Conf. on Color and Polarisation Techniques in Indust. Inspection, Germany, 1999.
5. F. Lumbreras et al. Color texture recognition through multiresolution features. Int. Conf. on Quality Control by Artificial Vision. 1:114–121, France, 2001.
6. J. A. Peñaranda, L. Briones and J. Florez. Color machine vision system for process control in ceramics industry. SPIE. 3101:182–192, 1997.
7. H. Kauppinen. Development of a color machine vision method for wood surface inspection. Phd Thesis, Oulu University, 1999.
8. J. Kyllönen and M. Pietikäinen Visual inspection of parquet slabs by combining color and texture. Proc. IAPR Workshop on Machine Vision Appl., Japan, 2000.
9. V. Lebrun and L. Macaire. Aspect inspection of marble tiles by color line-scan camera. Int. Conf. on Quality Control by Artificial Vision, France, 2001.
10. S. Kukkonen, H. Kvnen and J. Parkkinen. Color Features for Quality Control in Ceramic Tile Industry. Optical Engineering. 40(2):170–177, 2001.
11. R.O. Duda and P.E. Hart. Pattern classification and scene analysis. John Wiley and Sons, New York, 1973.
12. G. Wyszecki and W.S. Stiles. Color sciencie: concepts and methods, quantitative data and formulae. Wiley, 2nd Edition, New York, 1982.

Quantitative Identification of Marbles Aesthetical Features

Roberto Bruno[1], Lorenza Cuoghi[1], and Pascal Laurenge[2]

[1] Dipartimento di Ingegneria Chimica Mineraria e delle tecnologie Ambientali
University of Bologna, Bologna, Italy
[2] European Laboratory for Characterisation of Ornamental Stones, Bologna, Italy

Abstract. The use of image analysis for the aesthetical characterisation of stone slab surfaces has been studied during last ten years and has proved efficiency for an industrial and commercial application. This work aims to identify operational parameters specifically conceived for the classification of marble tiles. In this specific case the meaningful aesthetical properties are mainly linked to the anisotropy of the RGB intensities and, specifically, to the “veins”. Starting from the classical geostatistical and morphological modelling (variograms, granulometries, etc.), specific operational parameters have been obtained for a quantitative measurement of veins density, colour, and geometrical features (width, shape, continuity, etc.). The actual methodology defines commercial categories on the base of a self-appraisal process, which identifies intervals of several parameters. The current procedure is too rigid and doesn’t allow choosing in an intuitive way the discriminating properties. The proposed approach identifies understandable characteristics (vein features), and proposes quantitative indexes which actually satisfy the commercial classification of marbles.

1 Introduction

The economic value of the ornamental stones is strongly linked to their aesthetical properties. This work presents one of the possible methodologies which is able to characterise objectively the visual pattern of polished stone tiles and, more specifically, of the veined marbles. We will intend “veined” a stone which presents a high anisotropy of the spatial variability of visual properties. The quantitative characterisation of marble aesthetical features requires the application of specific instruments, different from the classical characterisation of images.

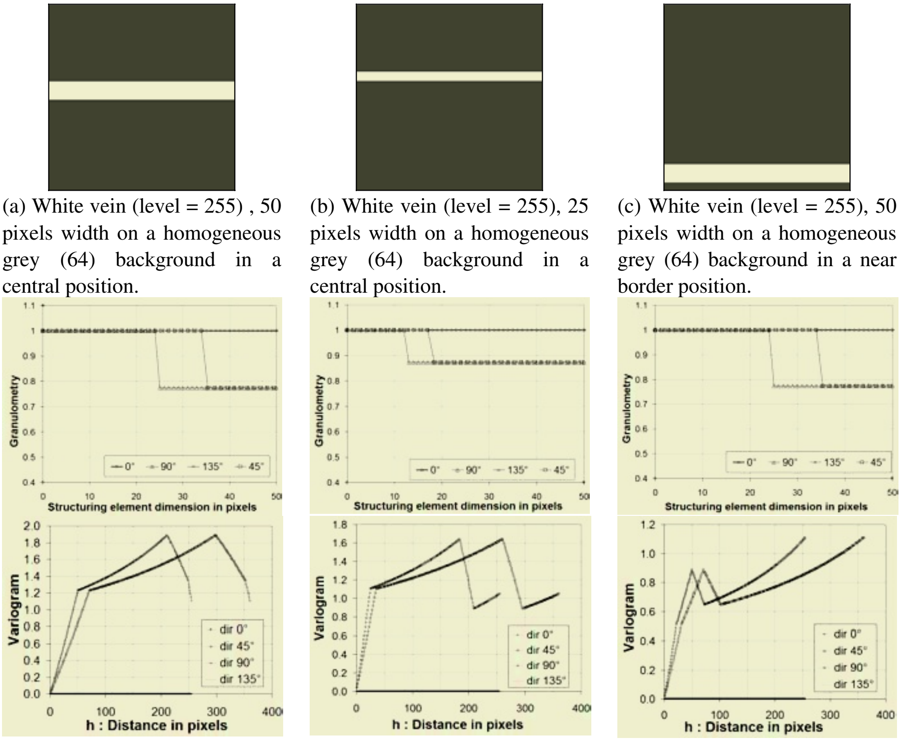
The presented methodologies are taken from the geostatistics and the mathematical morphology: specifically variogram and granulometry.

2 Overview of the Variograms and Granulometries Potentialities

We present now two methodologies that have the capacity to recognise and to measure the specific anisotropy of a tile’s image. To study the efficiency of these new instruments let us introduce simulated images which are a simplification of the real-

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 674–681, 2005.

ity. We resume the tile texture by an indicator variable, a homogeneous background with a “caricatured” vein. From the real image, we can obtain similar biphasic images by segmentation.



(a) White vein (level = 255) , 50 pixels width on a homogeneous grey (64) background in a central position.

(b) White vein (level = 255), 25 pixels width on a homogeneous grey (64) background in a central position.

(c) White vein (level = 255), 50 pixels width on a homogeneous grey (64) background in a near border position.

Fig. 1. Three examples of granulometries and variograms on synthetic images representing a simplified vein on a homogeneous background. These curves show what the granulometry and variogram can characterize in a veined tile image.

2.1 Asymmetric Granulometry

The function used for the granulometric curves is obtained applying iterative morphological openings. The based structuring element is an oriented line segment (3*1 pixel) with its origin at the centre. The opening is composed first by erosion and successively by dilation; thus the granulometry is useful for the characterization of bright veins on dark background (Fig. 1).

The asymmetric granulometry gives the following information (Fig. 1):

- the vein direction is given by the curve corresponding to the conservation after the successive openings of the original image, consequently where the granulometric curve equals to 1;
- the vein width is given by the smallest dimension (divided by 2) of the structuring element (x–axis) corresponding to the curve gap;
- the proportion of the veins on the image is given by the gap height (y-axis).

We should notice these results are independent of the vein position in the image field (except in the case the image border is included in the vein).

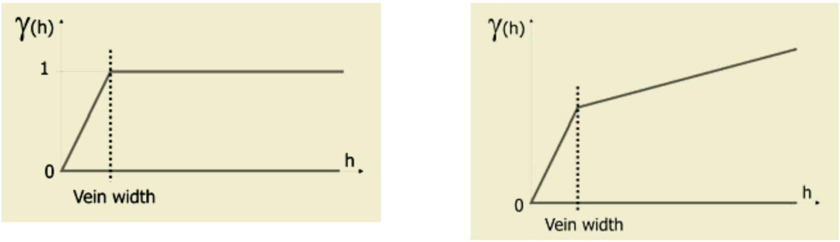
These measurements are always valid in the case of more veins on the image. From the values of the x-axis we can get the veins widths and, by intersecting the corresponding values of the gap proportion on the y-axis, we can get the veins number.

2.2 Variogram

We can consider the image pixels intensity as a realisation of a random function that the traditional geostatistics deal with.

If the variograms obtained are different in the various directions, we are in presence of anisotropy. The experimental variograms presented in the Fig. 2(a) are obtained in case of the perfect vein like the previous example but infinite. This behaviour corresponds to a zonal anisotropy, characterised by a sill equals to 0 in the direction of anisotropy.

The fact that we have to deal with a limited field (the image) produces a deviation respect to the ideal case, as we can notice in the Fig. 2(b).



(a) Variogram processed in an infinite field.

(b) Variogram in a limited field (the image).

Fig. 2. Expected variograms in function of the investigation field used to process it.

Referring to the Fig. 1, variograms give us more information respect to the granulometry. For example we can individuate the vein direction, that one with the variogram equals to 0.

The variogram is sensitive to the limited dimension of the image and to the position of the veins on it: the information we can get are concerning all the sides of the investigation field, not only the bright vein we want to characterise, but also the dark bands defined between it and the border.

It is possible to notice that the curve presents a change of slope or reaches a maximum or minimum point in correspondence, on the x-axis, of the widths of each band presented by the image. We can say that the significant point relative to the bright veins (giving, on the x-axis, the useful width) is the closest on the y-axis to the variance value, equal to 1 for a normalised variogram.

The reading of the variograms becomes more difficult at the increasing of veins number: all the changes of the intensity value in each direction are represented by a significant point on the curves.

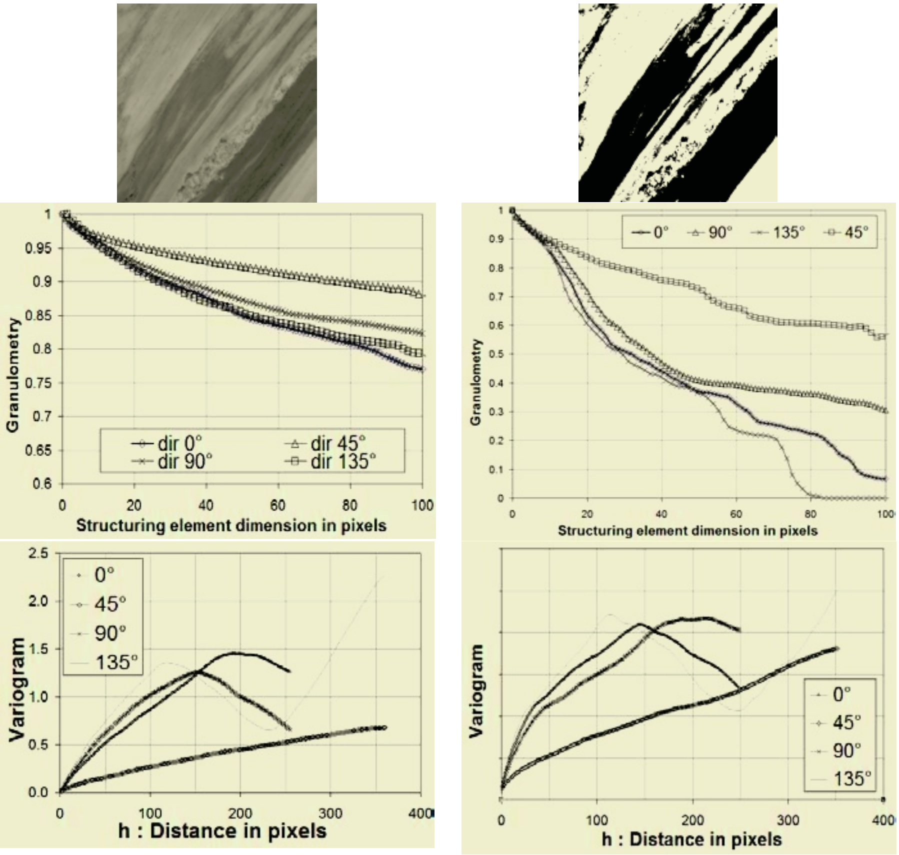


Fig. 3. Application of granulometry and variogram to the real tile image and to a segmentation of it.

2.3 Analysis of Results

In conclusion we can observe that the variograms give potentially more information than a first direct application of the asymmetric granulometry, but for this reason their application is very complicated.

The proposed tools consent to describe the veins on the images and to individuate their: direction, dimensions, percent on the total (and consequently number), position, complexity (linked to the gaps transformation in more or less complex continuity).

As we pointed out in the previous paragraph, the application of the two instruments for a univocal interpretation of the results can be very difficult because of the intersection of the several information. Fig. 3 shows an example of the behaviours of variograms and asymmetric granulometries in the case of a real tile image and of a segmented (binary) one.

The integrated use of the two instruments can be useful for resolving the complexity of reality.

3 Measurements Indexes

We propose some new parameters that link the advantage of a major strength of use to the possibility of supplying quantitative immediate valuation of aesthetical properties of tiles images.

To obtain the description of a marble sample, we can proceed step by step reducing the variability field of the material.

First we consider instruments by which it is possible to detect the presence or the absence of veins on the tile. The determining parameter to recognise a vein is essentially the colour; if this last one is uniform, so that there are not bands, but isotropic grains or spots, we can borrow the typical instruments used for the granites characterisation, as the segmentation.

If a given sample presents veined structures, as most of many marbles, the use of another specific instrument becomes necessary to distinguish the "not streaked" tiles, with a background easily recognisable, from the "streaked" tiles, for which there is a difficult interpretation of veins variability and background identification is not easy to obtain.

We proceed then in parallel to the identification of the methodologies allowing to read, veins characteristics for both aforesaid classes (streaked and not). More specifically, the following properties are required:

- dimensional (thickness and distribution),
- chromatic (colour passages vein - vein or vein - background),
- morphologic (edges and curving).

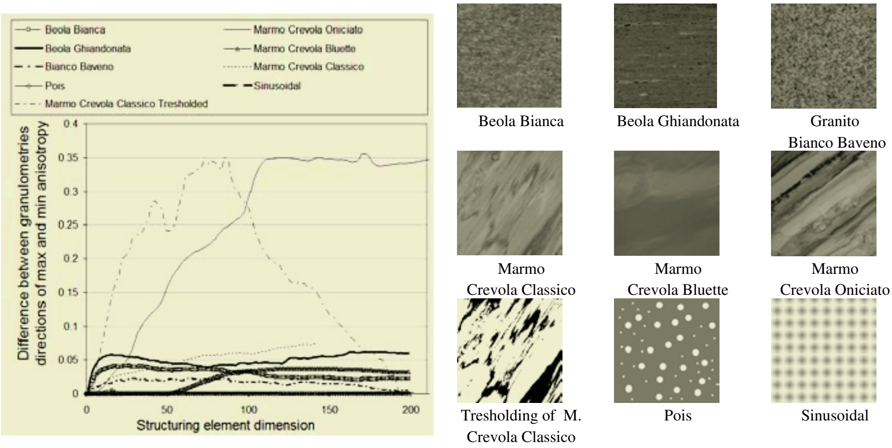
Finally, two diversified tools are necessary for the measurement of the chromatic properties in case of streaked or non-streaked tiles. Moreover, indexes are introduced which don’t request a segmentation process. In this work we present two examples of quantitative parameters useful for describing the anisotropy degree and the dimensional characteristics of the veins.

3.1 Vein Index

We consider “veined” a material that presents a significant degree of aesthetical anisotropy.

From the **granulometric** curves, whose effectiveness in the anisotropies measurements has been widely demonstrated, it is possible to determine some methods that give a quantitative measure of the anisotropy. The results are necessarily influenced by the chromatic contrast of the given material: the granulometry is a ratio between volumes, so that the anisotropy of the tiles, with veins strongly contrasted on the background, is easily recognisable. This behaviour, external from the objective of an anisotropy measure, reflects in some way the qualitative perception of the human eye: any material is more easily distinguished as “veined” if its veins are well recognisable rather than in the case of veins less contrasted with the background.

Among the several possibilities offered by the granulometric diagram to obtain an anisotropy measurement, we consider the difference between the more differentiated directional curves (Fig. 4).



Beola Bianca

Beola Ghiandonata

Granito
Bianco Baveno

Marmo
Crevola Classico

Marmo
Crevola Bluette

Marmo
Crevola Oniciato

Tresholding of M.
Crevola Classico

Pois

Sinusoidal

Fig. 4. Application of the vein index on natural and synthetic images.

The effect of the chromatic contrast is obvious in the shape of Marmo Crevola Oniciato and Marmo Crevola Classico Tresholded curves. The Granito Bianco Baveno presents a decreasing curve in the final part, which reflects the general isotropy of the material. The influence of the dimensions of the anisotropic elements in the determination of the index can be noticed: the Marmo Crevola Bluette is characterised by the presence of wide veins and lightly contrasted with the background. Its curve obtained by the chosen structural element (Fig. 4), till to the approximated dimension of 75 pixels, correspondent to the medium thickness of its bands, presents a behaviour more isotropic than the Granito Bianco Baveno, a typically isotropic material.

Table 1. Vein index classification corresponding to the images showed in the Fig. 4.

Image Name	*Vein index*	*Image Name*	*Vein index*
Uniform Grey	0	Marmo Crevola Bluette	0,0370
Pois	0	Beola Ghiandonata	0,0525
Sinusoidal	0	Marmo Crevola Classico	0,0694
Granito Bianco Baveno	0,0136	Marmo Crevola Oniciato	0,3494
Beola Bianca	0,024		

With the aim of getting for each image a unique value of reference and comparable among different materials, we choose as index the value supplied from the diagrams in correspondence of the abscissas of 125 pixels. Such threshold seems meaningful because for superior dimensions of the structuring element the curves present a plateau.

In the Table 1 are reported the values of the anisotropy index for the materials presented in Fig. 4. We have an anisotropy grading for the reported images. The results are quite interesting, because an anisotropy sill can be fixed, for instance equals to 0,03, and a classification of the images in two categories is obtained: the isotropic materials have an index smaller than 0,03, and the anisotropic ones higher. All the materials belonging to the second class present anisotropic structures, more or less thick and more or less contrasted with the background, and where other class distinctions can be done using these measurements.

An analogous vein index can be obtained taking advantage from the **variograms** curves; it is possible, for example, to analyse the ratio between the variogram values in the directions orthogonal and parallel to the actual veins.

Referring to the Fig. 5, we can observe that the presence of an anisotropy degree is associated to a lowering of the ratio curves. Uniform or isotropic materials, like for example the second tile of Marmo Crevola Bluette, present curves that remain constant on a value near 1.

A "U-shaped" behaviour, like for some tiles of Marmo Crevola Classico of the curves is linked to the presence of a geometrical anisotropy, so that it can be related to a complex structure of the investigated veins.

As we do for the granulometric index, we can empirically fix a limit value for distinguish veined and not veined samples: a tile that presents a curve becoming stable for a variograms ratio superior to 0,6 is completely isotropic.

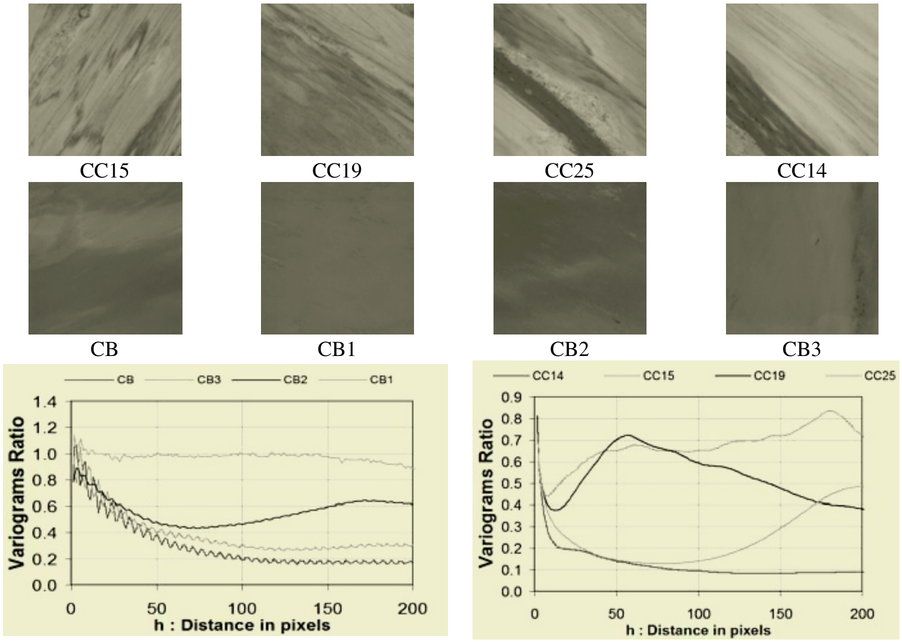


Fig. 5. Vein index processed by the variogram tool.

3.2 Dimensional Index

Using again the curves obtained by the application of asymmetric granulometry, we propose a numerical index that relates the veins dimension and the percent of the veins of a given dimension on the total tile image.

Let us consider "vein" any anisotropic structure, more or less continue, whose thickness does not exceed the dimensions of 200 pixels.

In Fig. 6 the requested steps to build the dimensional index are explained: the granulometric curve in direction perpendicular to the veins, that is the lower on the diagram, is considered; its value corresponding to the dimension of 200 pixels (i.e. to a structuring element dimension of 100 pixel) individuates the passed granulometric ratio; the dimension halving the "passing" ratio, identifies the median width of the vein.

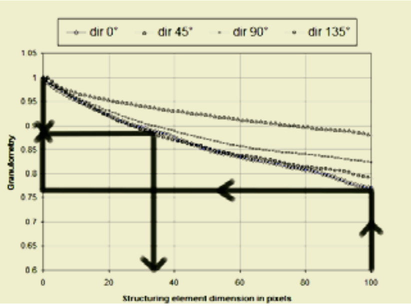


Fig. 6. Procedure to extract the dimension index on a granulometric curve.

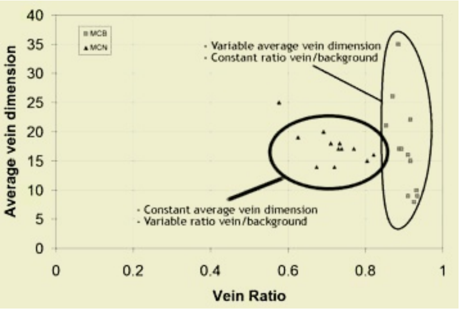


Fig. 7. Scatter plot between Vein dimension and vein index for two marbles presenting different kind of veins.

Referring to the Fig. 7 we can deduce some important information. A lengthened distribution of the points representing the samples of each material in horizontal or vertical direction on the diagram describes different veining characteristics, alternatively:

- a variable number of veins of similar dimension;
- a constant percent of differently dimensioned veins.

References

1. COSS, Characterisation of Ornamental Stone Standards by Image Analysis of Slab Surface, Final Report 1998, Contract N. SMT4-CT95-2028, DGXII E.C.
2. ELEFTHERIOU, N., 2001, Caratterizzazione Tecnico-Estetica delle rocce ornamentali coltivate nella provincia del Verbano-Cusio-Ossola, Eleftheriou, Bologna.
3. MATHERON G., 1965, Les variables régionaliées et leur estimation, Masson, Paris.
4. SOILLE P., 1999, Morphological Image Analysis, principle and Applications, Sprinter, Verlag.

Leather Inspection Based on Wavelets

João Luís Sobral

Departamento de Informática, Universidade do Minho
4710 - 057 Braga, Portugal
jls@di.uminho.pt

Abstract. This paper presents a new methodology to detect leather defects, based on the wavelet transform. The methodology uses a bank of optimised filters, where each filter is tuned to one defect type. Filter shape and wavelet subband are selected based the maximisation of the ratio between features values on defect regions and on normal regions. The proposed methodology can detect defects even when small features variations are present, which are not detect by generic texture classification techniques, and is fast enough to be used for real-time leather inspection.

1 Introduction

Leather inspection has been recognised as one very complex problem on the area of texture classification. Like most of natural textures, feature values have a high variation, forming a pseudo-random structure, e.g., features present a high variation but still follow a statistical distribution. Machine made textures have a much more regular pattern, since features values follow a more predefined range.

Generic texture classification methods assign a texture to one of a set of predefined classes. Typical texture features for each texture class are computed from texture samples. Comparing the texture features with the features from each texture class, using some proximity criterion, can classify an unknown texture.

Defect detection is usually based on a simpler texture classification scheme, using just a defect class and a normal class. The classification can be based on a threshold classifier that assigns a texture to one of these two classes. Unfortunately, the difference among features in defect types can be large, leading to several defect classes and to a more complex classifier. Also, features difference from normal and defective regions can be very small, leading to a poor performance when using generic texture classification schemes.

This paper presents a methodology to select a set of optimised filters, where each filter is applied to the texture wavelet transform. The process searches for filters that achieve the highest ratio between features on normal and defective regions for each defect type. Each filter is based on a wavelet sub-band and can have different size and orientation.

The rest of this paper is organised as follows. Section 2 presents an overview of related work, mainly based on filter bank and wavelet based approaches. Section 3 introduces the methodology and section 4 presents performance results. Section 5 closes the paper with suggestions for future work.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 682–688, 2005.

2 Related Work

The filter bank approach is one of the most used approaches for texture classification. It is based on a bank of convolution masks to extract texture features. Gabor filters [1] have been used for texture classification because of their optimal space/frequency. Wavelet transforms [2] are very efficient to calculate and have also been used for texture classification [3][4]. A comparative study of filter bank approaches for texture classification can be found in [5].

Optimised techniques for texture classification are based on filters that maximise the features difference among textures. In [6] a method is presented to design an optimal finite impulse filter for defect detection in textured materials and [7] uses a combination of Wavelets and Co-Occurrence Matrices. Optimised Gabor filters also have been used in [8][9].

3 Proposed Methodology

This section begins by presenting wavelet-based methodologies for texture classification, then the defect detection methodology is presented and finally the process of selecting the optimised filters is analysed.

3.1 Wavelet Texture Classification

Texture classification based on wavelet transform is based on a successive application of a low pass and a high pass filter and on sub-sampling after each filter application. There are several wavelets, each one uses different low pass and high pass filters. The simplest and fastest wavelet is the Haar Wavelet, which has a low pass filter equal to [0.5, 0.5] and the high pass filter equal to [-0.5, +0.5].

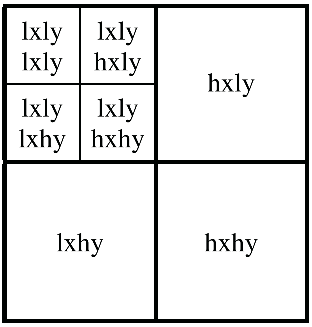


Fig. 1. Sub-band representation.

Figure 1 shows a sub-band representation of a two level wavelet transform and the sequence of filters applied on each sub-band. For example, the top right sub-band is obtained applying a high pass filter in the horizontal direction (hx), sub-sampling to retain half of the pixels, applying a low pass filter in the vertical direction (ly) and sub-sampling again. Texture classification can be performed using the average energy on each sub-band as texture features. In this example there are 7 features per texture.

3.2 Defect Detection

Leather defects can have several shapes and sizes depending on the defect type. Namely, there are small defects with a high feature difference from normal areas and there are large defects that have a small feature difference from normal areas. This means that using a single filter size does not achieve a good performance: small defects require small filter sizes and large defects require larger filter sizes. Also, defect shapes are not the same for all defect types, some defects occupy a long and tiny area while others occupy a small or large square area.

Traditional filter bank approaches (Figure 2) apply a bank of filters to the input image, followed by a nonlinear function (usually squaring) and followed by a smoothing filter (averaging). The smoothing filter size is usually calculated based on the central band-pass frequency of each filter. A classifier, assigning each texture to a class, uses the resulting feature vectors from the smoothing step.

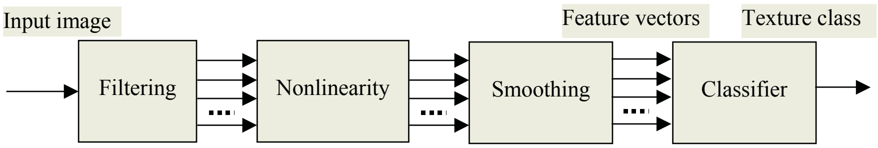


Fig. 2. Filter bank based texture classification.

The proposed methodology uses a filtering step based on wavelet packets, which achieves the sub-band decomposition presented in Figure 3. This decomposition uses both the first and the second level of decomposition and generates 20 (4 + 16) features per texture.

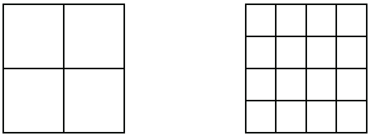


Fig. 3. Wavelet sub-band decomposition.

The nonlinear processing step is not applied in this methodology, allowing the measurement of both positive and negative filter responses, improving the detection for some type of defects.

The smoothing step is based on a set of filters, each filter having a size and shape tuned to a defect type. Figure 4 presents an overview of the recognition process.

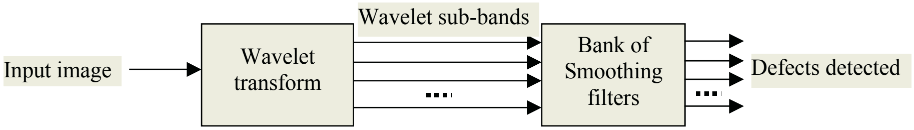


Fig. 4. Defect detection process.

The wavelet transform produces 20 feature images, each filter is then applied to one of these feature images and its output will signal the detection of one defect type.

Note that several filters may detect a particular defect, namely, on defects that have a high feature difference from normal regions.

3.3 Filter Selection

Each smoothing filter is tuned for a kind of defect. The output from an optimal filter should maximise the ratio between defect regions and normal areas. The process of selecting the best filter is based on training samples with manually identified defects.

Optimised filter selection is performed by first segmenting the image into a number of small regions. A set of predefined smoothing filters is then applied to each region and to each wavelet sub-band, obtaining a large number of features per region. The best filter is afterwards selected as the one that provides the maximum feature ratio between each defect region and the average of other regions in the same sample.

The smoothing filters perform pixel averaging over a 3x3 to 13x13 pixel neighbourhoods, both square and rectangular (e.g., 3x5, 3x7, …). Also, for rectangular neighbourhoods, a 45° rotated mask is tested. These rotated masks are more appropriated to detect thin defects that are not horizontal or vertical. Overall the process tests 2 x 36 filters mask, each one applied to all the wavelet sub-bands. This leads to a total of 2 x 36 x 20 (e.g., 1440) possible masks, which is considerable less than other techniques [8][9], where the optimised filters have several free parameters, leading to a complex scheme to find the best filter parameters.

Smoothing filters can be viewed as a frequency subdivision of each wavelet sub-band, but since we tested several rectangular shapes and orientations the subdivision is not restricted to a square frequency notch. Also, this process tunes the filters to detect missing frequencies or positive/negative ramps (e.g. local frequencies with a high amplitude), and their size matches the expected minimum defect area. This is achieved by using minimum or maximum filter response on the defect area instead of the usual average filter response. Using minimum or maximum filter response also helps to overcome the difficulty of the accurate image segmentation.

Overall the process are based on the wavelet transform to perform a fast square partition of the frequency spectrum, followed by a finer partition, based on smoothing filters. Also note that since both the first and second levels of the wavelet transform are used, it is possible to detect defects that affect a large area of the frequency spectrum (first level), as well as defects that affect a smaller area (second level).

We found by experimental work that a ratio larger than six ensures a detect detection free from false positives (e.g., regions detected as defects that are not defects). If no filter response achieves this ratio it means that the defect cannot be detected using this kind of filters.

Each defect sample usually leads to a specific filter, so a second step selects the most generic filters. In this step the filters are successively selected based on their ability to detect the highest number of defects. This process successively selects the filter that adds the largest number of defects that are not already detected by the previously selected filters. To avoid too much filter specialisation (e.g., each filter detecting just one defect), each optimised filter is required to detect several defects.

4 Results

The proposed methodology has been applied to several leather samples. Each sample was taken with a resolution of five pixels per millimetre. The image brightness and contrast are normalised to eliminate illumination effects. It is then segmented to obtain a large amount of regions. This segmentation step selects the largest regions, which have a low brightness or a low or high contrast, which always includes defect regions and a high amount of other regions. Both filter selection (e.g., training) and leather inspection is performed using 200 regions per 4000x2000 pixels area.

We applied the proposed methodology to a database of about 150 samples. Using the undecimated Haar Wavelet transform and 8 optimised filters, it was able to achieve the same recognition rate as an experienced human operator. Filter selection was performed using another database of 20 defects. Note that many of the defects require a highly trained operator to be manually detected, since there is only a slight feature variation on the defect area, and that most of these defects are not detected by more traditional methodologies, such as Gabor filters or normal Wavelet Packets.

Figure 6 presents three small samples; the first one is detected by traditional approaches, while the second and the third are not detected. Figure 5 presents the highest ratio between the defect region and normal regions obtained on each wavelet sub-band, respectively, on the second and on the third leather sample. On both these samples the maximum ratio was obtained using a 45° smoothing 7x3 filter. Note that some defects can be detected on several sub-bands. In such case, usually the lower level is preferred, since it involves a lower number of calculations. Also note that, in these two cases, using a square mask (or a non-rotated) does not achieve a ratio high enough to provide a reliable detection.

4.2	1.7
5.4	1.4

6.7	2.5
5.3	0.2

4.4	1.7	2.3	3.0
6.1	3.0	2.1	5.3
5.9	2.4	1.5	3.1
4.3	5.1	3.3	2.7

6.6	3.5	3.9	0.7
5.6	1.7	1.7	1.0
5.6	1.9	0.8	0.1
3.3	0.6	0.6	0.7

Fig. 5. Maximum ratio defect / normal region per wavelet sub-band.

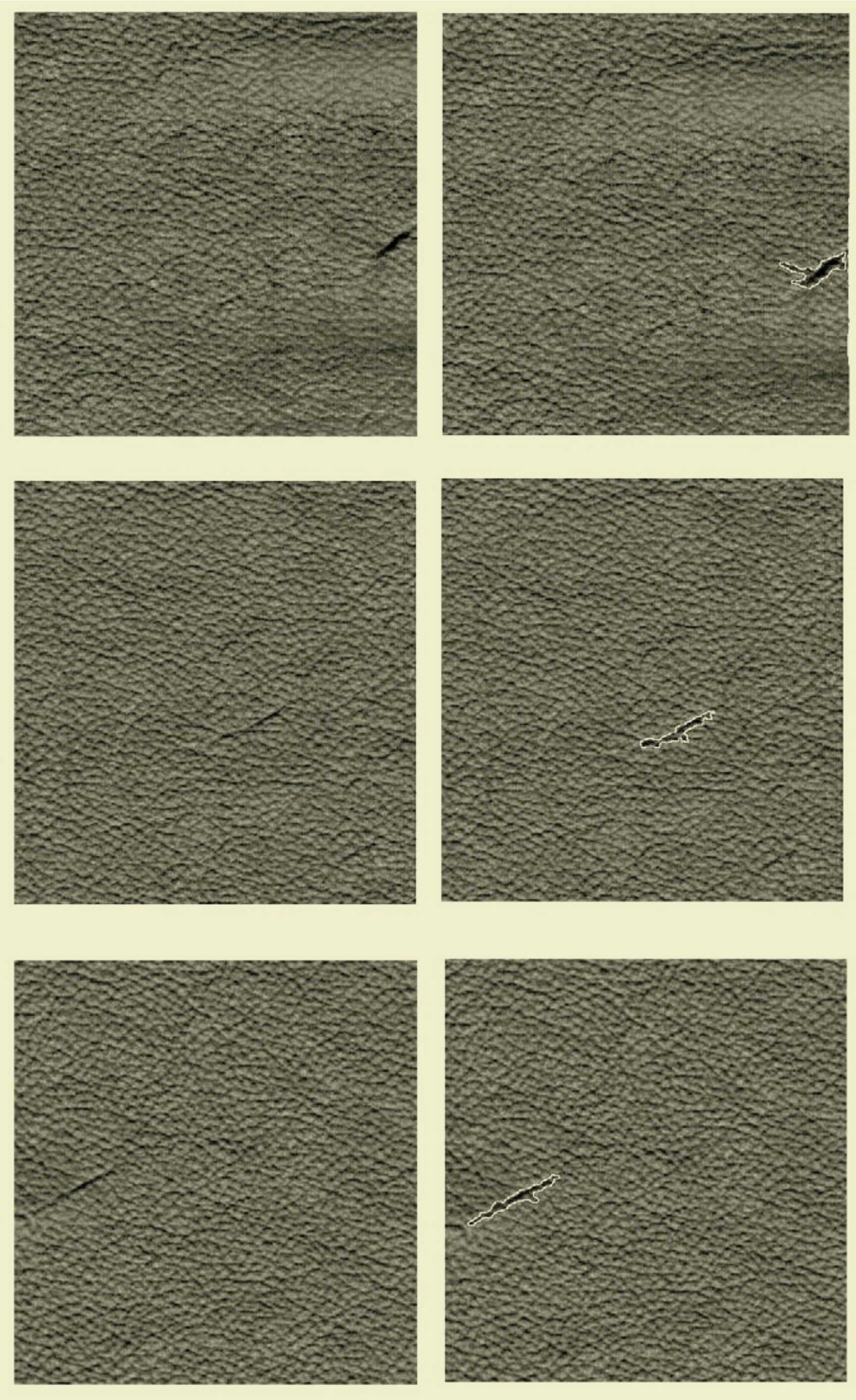


Fig. 6. Leather samples and inspection results.

In spite of having a *similar* look, the third defect is larger than the second, so it requires a different mask and can be detected on the first level, in the lxly sub-band. The second defect can be detected in the lxlylxhy or in the lxhylxly sub-band.

The inspection process takes 70 seconds for an image of 4000x6000 pixels (leather size of 0.8m x 1.2m), which is less than the time required for an automatic machine to cut the leather. These times were collected a PC AMD XP1800+, running Red Hat Linux 8.0 and the GNU Gcc 3.2 compiler.

5 Conclusion

This paper presented a methodology for leather inspection. The methodology is based on a wavelet transform and on a bank of optimised smoothing filters, each filter tuned for a particular defect type.

The methodology was able to detect several types of defects that current approaches cannot detect, it achieves the same detection rate as an experienced operator and it is fast enough to be used for real time inspection.

This methodology was applied to leather inspection but we think that it can be also applied to other inspection areas.

Current research includes using more features per optimised filter, since each filter is based on a single wavelet sub-band. This improvement should increase the recognition rate but it will also increase the complexity of the process. Ongoing work also includes experiments with other Wavelet Transforms and Gaussian smoothing filters. However all these enhancements will increase the processing time so other improvements are required to achieve the same processing time.

References

1. Jain, A., Farrokhnia, F.: Unsupervised Texture Segmentation Using Gabor Filters, IEEE International Conference on Systems, Man and Cybernetics, Nov. (1990)
2. Mallat, S.: A Theory for Multiresolution Signal Decomposition: The Wavelet Representation, IEEE Trans. Pattern Analysis and Machine Int., vol. 11(7), July (1989).
3. Laine, A., Fan, J.: Texture Classification by Wavelet Packet Signatures, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15(11), Nov. (1993)
4. Unser, M.: Texture Classification and Segmentation Using Wavelet Frames IEEE Transaction on Image Processing, vol. 4(11), November (1995)
5. Randen, T., Husoy, J.: Filtering for Texture Classification: A Comparative Study, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21(4), April (1999)
6. Kumar, A., Pang, G.: Defect Detection in Textured Materials Using Optimised Filters, IEEE Trans. Systems, Man and Cyber. – Part B: Cybernetics, vol. 32(5), October (2002)
7. Amet, A., Ertuzun, A., Ercil, A.: Texture Defect Detection using Sub-band Domain Co-Occurrence Matrices, IEEE Southwest Symposium on Image Analysis and Interpretation, April (1998)
8. Kumar, A., Pang, G.: Defect Detection in Textured Materials Using Gabor Filters, IEEE Transactions on Industry Applications, vol. 38(2), March/April (2002)
9. Bodnarova, A., Bennamoun, M., Latham, J.: A Constrained Minimisation Approach to Optimise Gabor Filters for Detecting Flaws in Woven Textiles, IEEE International Conference on Acoustics, Speech, and Signal Processing, June (2000)

Multispectral Image Segmentation by Energy Minimization for Fruit Quality Estimation*

Adolfo Martínez-Usó, Filiberto Pla, and Pedro García-Sevilla

Dept. Lenguajes y Sistemas Informáticos, Jaume I Univerisity
{auso,pla,pgarcia}@uji.es
http://www.vision.uji.es

Abstract. This article presents the results of an unsupervised segmentation algorithm in multispectral images. The algorithm uses a minimization function which takes into account each band intensity information together with global edge criterion. Due to the unsupervised nature of the procedure, it can adapt itself to the huge variability of intensities and shapes of the image regions. Results shows the effectiveness of the method in multispectral fruit inspection applications and in remote sensing tasks.

1 Introduction

The main motivation of the developed work has been to obtain a method able to segment images of fruits for their quality classification in visual inspection processes using multispectral information. Particularly, this application problem implies the following requirements:

1. The method has to be able to segment *multispectral images* obtained from several wavelength ranges.
2. An *unsupervised method* would be needed due to manifold variables which can arise in fruit images. Thus, any prior knowledge should be avoided for the segmentation procedure.
3. The segmentation method has to be mainly based on *multispectral intensity* and *edge criteria*, in order to define the segmented region boundaries as accurately as possible.

The traditional effectiveness of a multiresolution Quadtree (QT) structure is combined with the multispectral intensity and edge information in a hierarchical representation. This leads us to an efficient strategy to solve the problem.

The method we are presenting starts from [4] as a natural development to multispectral images. Although this algorithm has achieved good results, it is quite obvious that multispectral images may offer us a lot of new and useful information. More concretely, in fruit inspection tasks we realize that there exist defects that can only be detected in certain bands of the spectrum and most of

* This work has been partly supported by grants DPI2001-2956-C02-02 from Spanish CICYT and IST-2001-37306 from the European Union

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 689–696, 2005.

the defects have a specific range of bands where they can be better discriminated. Therefore, fruit inspection by means of multispectral images acquires a high importance when we want to increase the classification quality and to obtain more accurate results in our application.

Multispectral images involve larger amount of information, noise, etc., so a preprocessing step has been included in order to improve the final results and accelerate the whole process:

1. On the one hand, we use an invariant representations of the input values to obtain multispectral images independent from some lighting and geometrical conditions [5]. These representations are very useful in those tasks that require invariance to shadows, orientation, reflections, etc.
2. On the other hand, we accelerate the whole process by means of the band selection proposed in [7]. Together with this acceleration, we have been able to keep the quality of the segmentation results just like if no band selection would have been applied.

The main contribution of the presented work is the proposed multispectral energy function that efficiently combines intra-region features with border information. The proposed framework yields satisfactory results, particularly in fruit inspection tasks.

2 Variational Image Segmentation

Variational methods for image segmentation develop algorithms and their mathematical analysis to minimize the segmentation energy E represented by a real value. The segmentation energy measures how smooth the regions are, the similarity between the segmented image and the original one and the similarity between the obtained edges and the discontinuities of the original image.

The Mumford-Shah model [6] has been regarded as a general model within variational segmentation methods. According to Mumford-Shah's conjecture, the minimal segmentation exists but it is not unique; for each image a set of minimal segmentations exists. This model looks for a piecewise smoothed image u with a set of discontinuities, edges of the original image g by means of minimizing the segmentation energy in the following equation, where K is the set of discontinuities in the image domain Ω representing the *edges* of g:

$$E(u,K) = \int_{\frac{\Omega}{K}} (|\nabla u(x)|^2 + (u-g)^2)dx + length(K) \tag{1}$$

Since Mumford-Shah's work, several approaches appeared that suggested modifications to the original scheme. Recent works change equation (1) in order to improve the results. In this sense, the boundary function, which is binary in the Mumford and Shah's formulation, was changed by a continuous one which obtains a clearly defined boundary in [2]. Furthermore, in [1] the authors analyze some possible generalizations of the Mumford-Shah functional for color images. They suggest that these changes accentuate different features in edge detection and restoration.

In general, formulating variational methods have several advantages:

1. A variational approach returns explicitly a measure of the quality of the segmentation. Therefore, on the one hand we are able to know how good the segmentation is, on the other we may use it as a quantitative criterion to measure the segmentation quality.
2. Many segmentation techniques can be formulated as a variational method.
3. Finally, a variational approach provides a way to implement non-supervised processes by looking for a minimum in the segmentation energy.

3 Energy Minimization of the Quadtree Image Representation

In this work, a function to minimize the segmentation energy is proposed. With this assumption, it is important to point out that we cannot guarantee to find a global minimum but, the experimental results obtained show that the solutions are very satisfactory. In addition, we use a statistically robust functional that takes into account any resolution or scale change producing the same segmentation results in each case.

Let u be a smoothed image and a set of discontinuities of the original image g, let R_i be a set, with $0 < i \leq \|\Omega\|$ and $R_i \subseteq \Omega$. R_i is a family of r regions in Ω such that $\bigcup_{i=1}^{r} R_i = \Omega$ and $R_i \bigcap R_j = \emptyset$ for $i \neq j$. Let B_i represent the border of region R_i, that is, $R_i' = R_i - B_i$ is the inner part of region R_i. Finally, let γ be certain very small value that avoids dividing by zero.

Thus, $\forall x \in R_i$ let us consider the function $E_i(R_i, B_i)$:

$$E_i(R_i, B_i) = \int_{R_i} (|u(x) - m_i|)\ dx + \frac{\int_{R_i'} |\nabla g(x)|}{\int_{B_i} |\nabla g(x)| + \gamma}\ dx \tag{2}$$

In the image u with N bands, $u(x) = (u_1(x), ..., u_N(x))$ represents the intensity values in each band $u_j(x)$ of an R_i element x, and m_i is a central measure for the intensity value of R_i. The final segmentation energy is expressed as $E(\Omega) = \sum_i E_i(R_i, B_i) + \|\Omega\|$.

The QT structure allows us to divide an image within a complete multiresolution tree representation including neighboring information. This spatial information can be further used by a clustering strategy which groups the QT leaves using the intensity values in each band and the edge information.

Let us see the following discrete version of (2) with the same nomenclature,

$$E_i(g) = k \cdot H_i + (1 - k) \cdot G_i + \lambda \cdot length(\Omega) \tag{3}$$

that returns the energy at each region. H_i and G_i are terms as follows:

$$H_i = \frac{\sum_{R_i}\ D(u(x), m_i)}{\sigma_{image}} \qquad\qquad G_i = \frac{\sum_{R_i - B_i} |\nabla g(x)|}{\sum_{B_i} |\nabla g(x)| + \gamma} \tag{4}$$

Specifically, in the QT representation, R_i is the set of leaves of the QT belonging to region i and B_i represents the boundary leaves in R_i, $0 < i \leq r$, with r being the number of regions at each iteration. The function D calculates a distance between two intensity sets of pixels (Euclidean, Manhattan, etc.). The value $|\nabla g(x)|$ returns the gradient magnitude at the pixel x. Note that the parameter $0 < k < 1$ allows to define the weight between the color and boundary information and λ allows the method to give more or less importance to the number of regions. Finally, the value σ_{image}, which is the sum of the standard deviations of each plane in the image, contributes to normalize the first term and makes the function statistically robust. Thus, in the energy functional (3) we can distinguish three components:

1. The first one takes into account the homogeneity of each region by means of a distance from each pixel to a central measure of the region it belongs (usually the median). So, the smaller this result, the more homogeneous the cluster is. This component represents our specific constraint to each region with the intensity information in each spectral band.
2. The second component promotes that the gradient magnitude will be low in the inner leaves and high in the boundary ones, that is, promotes regions with no edges inside. In this sense, gradient information is used by means of a boundary map which is made from the maximum gradient value found in each band.
3. The third term helps the function to punish a large number of regions and it is frequently used in variational image segmentation as in [2][1].

It is important to point out that the energy functional (3) has been designed in order to achieve a statistically robust behavior, being invariant to scale or/and intensity changes.

4 The Algorithm

4.1 The Preprocessing Step

1. The collection of multispectral images were obtained by an imaging spectrograph (RetigaEx, Opto-knowledged Systems Inc., Canada). The spectral range extended from 400 to 720 nm in the visible, with a spectral resolution of 10 nm, obtaining a set of 33 spectral bands for each image.
2. Multispectral applications with a limited time to finish usually have to deal with the band selection problem in order to characterize the problem without loss of accuracy. In this sense, considering that multispectral images contain information represented by means of a set of two dimensional signals, in [7], it is proposed a band selection from the point of view of information theory measures. So, dealing with the relationships among the set of bands representing the multispectral image, the optimal subset of spectral image bands can be obtained.
   Using a database of oranges, in [7] is shown that a 97.4% of classification performance is achieved with 6 bands, whereas using th 33 bands they manage a 98.5% in their unsupervised classification approach. We use the same database and select the same 6 resulting bands as in [7].

3. In [5], authors presented a set of illumination invariants to work with multispectral images. The results obtained showed the good performance of those invariants in order to avoid changes in the illumination intensity, highlights and object geometry. Thus, after the band selection, we apply the transformation:
$$C_n^{new} = \frac{C_n - \min(C_1, \ldots, C_N)}{\sum_j (C_j - min(C_1, \ldots, C_N))} \tag{5}$$
to obtain an invariant representation where C_i is the pixel value in band i and $1 \leq i \leq N$.

4.2 The Segmentation Process

The QT structure allows us to divide an image within a complete multiresolution tree representation including neighboring information. This spatial information can be further used by a clustering strategy which joins the QT leaves using intensity and edge information. The multiresolution process allows us to analyze the image with a coarse-to-fine sequence described as follows:

1. We construct a hierarchical structure level by level. It is important to clarify that, talking about a QT, the meaning of *level* is a set of leaves with the same size. Thus, the first level will be the first four children leaves that descend from the whole image and so on. Therefore, while the previous named QT is created, each level is revised by the functional (3) in order to revise the clustering at that resolution. Each cluster created in any level will be taken into account in the next levels. Finally, when we finish the QT construction, the *salient regions* have been detected in a coarse way.
2. Focusing the attention on the *salient regions* (the coarse ones that have been labeled), they will be taken as the most significant groupings of the image. So, we continue the process expanding each cluster by means of a region growing method where each cluster applies the functional (3) to its neighboring regions. This second step will take care of shaping the edges of each region by intensity and edge criteria.

Note that we use the functional (3) described in Sect. 3 in both of the previous steps but, whereas in the first one this application is guided by a hierarchical structure in order to develop each resolution level, in the second one the application of the functional follows a region growing strategy to achieve the final regions in a more accurate way.

Before summarizing the segmentation process, it is important to point out what are the main ideas the proposed method is based on:

1. We look for different features provided by the invariants in a first step. In addition, we accelerate the process by selecting an optimal set of spectral bands without loss of quality.
2. To guide the segmentation process, the following questions have to be solved:
   (a) *the way to continue the segmentation process.*
   (b) *how long the process have to continue.*

The first question is solved by means of a multiresolution analysis of the image with a QT structure. Multiresolution is able to decompose the image in several resolution levels developing a coarse-to-fine process from the *salient regions* to the final shapes of each region. On the other hand, the question (b) will be determined by the functional (3) described in Sect. 3. It will be minimized in a progressive way until the functional energy stops decreasing.

The whole segmentation process is summarized in the following iterative algorithm where each cluster is compared to all its neighboring clusters and it will be merged when the segmentation criterion is satisfied (see Sect.3).

1. Apply the band selection (6 from 33) and change the representation to the one described in [5].
2. Make the boundary map as an edge information reference.
3. Construct an oversegmented representation of the image, that is, expand the QT until every square region have all pixels with the same intensity. After this, create an ordered list according to region sizes.
4. The algorithm computes the functional (3) for each region and its neighbor regions in an iterative sequence that may be seen as a coarse-to-fine segmentation process.
5. If the whole image energy has decreased, reorder the list of regions by size and repeat the previous step. Arranging regions according to their size gives more importance to bigger regions and represents a spatial constraint that facilitates merging small regions with big ones.
6. Ignore very small regions (not identifiable with any defect) and merge them to their most similar neighbors.

This clustering process stops when no other merging can be performed without increasing the energy of the segmentation.

5 Results

Fruit image segmentation is used as input in a further process to characterize and classify the fruit surface. These visual inspection applications identify and detect different types of defects and parts of the fruit.

Fig.1 and Fig.2 show examples of oranges with two different types of defects on their surface, overripe defect and scratch defect respectively. In both images, the first row shows some image bands. These bands are not the ones selected by the band selection method, but they were chosen for illustration purposes.

Second row shows the edge results, where the darker the edge line is, the greater difference between the spectral values of neighboring regions. The first image on the left is the result after the segmentation algorithm is applied over all 33 bands. As we can see, this segmentation tends to finish with too many regions due to illumination and geometrical factors. When the segmentation algorithm is applied on the invariant representation (second images on the second rows) we obtain satisfactory results. In the last two images where the band selection and the invariant representation were used, the quality of the results increased, since

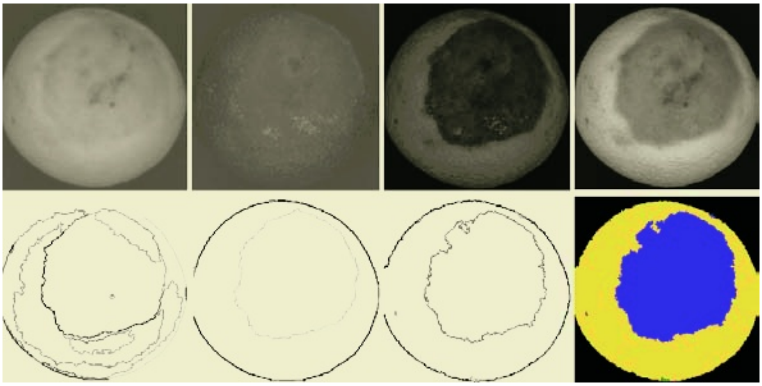


Fig. 1. First row corresponds to the bands 0, 8, 16 and 32 of an orange with an overripe defect in its surface (that is, 400, 480, 560 and 720 nm in the visible spectrum). Second row shows, from left to right, results using 33 bands, illumination invariants, 6 bands and illumination invariants and the final segmentation represented with random colors.

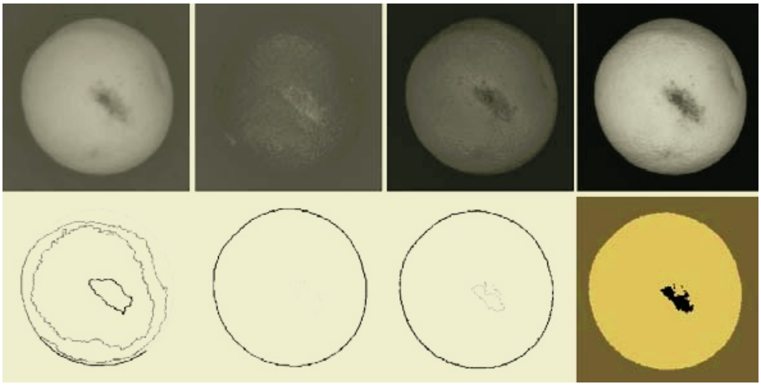


Fig. 2. As Fig.1, first row corresponds to the bands 0, 8, 16 and 32, but this orange has a scratch defect in its surface. Second row shows the edge results and the final segmentation in the same order as Fig.1.

the boundaries are more accurately defined and some new regions are discovered. Moreover, note how the segmentation process adapts to the regions of each image due to its unsupervised nature.

Thus, the segmentation obtained has found the different variations of the stains on the orange surface and this will allow the extraction of region descriptors for their classification in fruit quality assessment applications.

In order to show the performance of the presented method to other application fields, Fig.3 shows a multispectral image from satellite and the results obtained. This image has 220 bands and some of them are quite noisy (the special features of this image are explained in [3]). As we can see, the final segmentation has separated the brighter regions and the darker ones drawing the boundaries quite accurately.

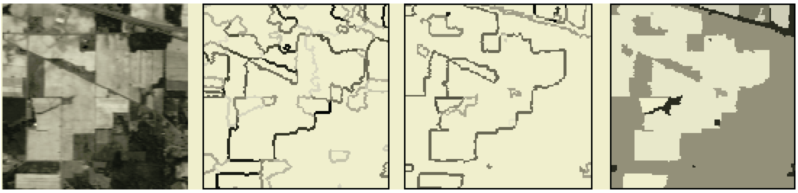


Fig. 3. Satellite image results. From left to right, original image (band 120), results using the band selection, results using the same selected bands and illumination invariants and the final segmentation represented with random colors.

6 Conclusions

In this paper, we have presented a multispectral segmentation algorithm based on a minimization function that uses the intensity and edge information. This algorithm has been combined with a preprocessing step which improves and accelerates the whole process. The results obtained show how the algorithm can adapt to the different situations and variability of intensity regions, being able to segment areas and locating the borders due to the use of gradient information during the segmentation process. Thus, this unsupervised segmentation strategy can locate different regions and find their contours satisfactorily.

Enclose to the minimization process we have developed a preprocessing step, based on a band selection and illumination invariants for multispectral images, which improves notably the accuracy of the results and a QT representation that guides the minimization process and increases the efficiency by means of a segmentation at different resolution levels.

References

1. A. Brook, R. Kimmel, and N.A. Sochen. Variational restoration and edge detection for color images. *Journal of Mathematical Imaging and Vision*, 18(3):247–268, 2003.
2. G.A. Hewer, C. Kenney, and B.S. Manjunath. Variational image segmentation using boundary functions. *IEEE Transactions on Image Processing*, 7(9):1269–1282, 1998.
3. Luis O. Jimenez and A. Landgrebe. Hyperspectral data analysis and supervised feature reduction via projection pursuit. *IEEE TGRS*, 37(6):2653–2667, 1999.
4. A. Martinez-Uso, Filiberto Pla, and Pedro Garcia-Sevilla. Color image segmentation using energy minimization on a quadtree representation. *LNCS*, (3211):25–32, 2004.
5. R. Montoliu, F. Pla, and Arnoud K. Klaren. Multispectral invariants. Technical Report, DLSI, Universitat Jaume I, Castellon, Spain, 2004.
6. D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *CPAM*, 42(4), 1989.
7. Jose M. Sotoca, Filiberto Pla, and Arnoud K. Klaren. Unsupervised band selection for multispectral images using information theory. *ICPR*, 2004.

Thresholding Methods on MRI to Evaluate Intramuscular Fat Level on Iberian Ham

Mar Ávila[1], Marisa Luisa Durán[1], Andres Caro[1],
Teresa Antequera[2], and Ramiro Gallardo[3]

[1] University of Extremadura, Computer Science Dept.
Escuela Politécnica, Av. Universidad s/n, 10071 Cáceres, Spain
{mmavila,mlduran,andresc}@unex.es
[2] University of Extremadura, Food Technology
Facultad Veterinaria, Av. Universidad s/n, 10071 Cáceres, Spain
tantero@unex.es
[3] "Infanta Cristina" University Hospital, Radiology Service, Badajoz, Spain

Abstract. Thresholding techniques are the simplest and most widely used methods to automatically segments images. These are used to segment images into several regions. This paper works over two sets of Iberian ham images: images taken by a digital camera (CCD) and Magnetic Resonance images (MRI), in order to establish a comparative for the performance on each kind of images. A methodology to determine the intramuscular fat (IMF) level of Iberian ham using computer vision techniques has been developed, as an attempt to find an alternative methodology to the traditional and destructive methods. The correlation between the chemical data and the computer vision results have been established in the paper. The main conclusions of the work are that better results have been obtained for MRI, which do not require preprocessing methods. So, the proposed approach to determine the IMF level could be considered as an alternative to the traditional and destructive methods.

1 Introduction

In the last years, several Computer Vision techniques have been applied on images in the field of Food Technology, in an attempt to find alternatives to the traditional, and generally destructive, methodologies.

Different studies have proved that juiciness and acceptance of the meat is mainly influenced by its fat content, either in calf [13], pig [2] or lamb [20] meat. This work is centered on Iberian pig. Its meat is mainly used to produce dry-cured products, in particular, cured hams. The special features of the Iberian pig meat (high intramuscular fat content, fatty acid composition, antioxidative status...) together with the prolonged ripening process produce a dry-cured ham with a quite special flavor that makes it one of the most valuable meat products by his high quality.

In relation with the fat content in Iberian pig ham, chemical analysis is the most frequently used method to determine the intramuscular fat (IMF) level [1,17]. But, this technique is expensive, destructive and tedious.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 697–704, 2005.

In previous works Iberian ham images have been processed for different purposes [3, 4, 7, 8]. The feasibility of alternative techniques to determine IMF level using Computer Vision techniques has been studied in [7, 8, 14]. In [14] the IMF level was determined from images of Iberian raw meat of thigh muscles, which were cut in slices and digitized by a CCD camera. These digital images were segmented in two regions (fat and background) using a semiautomatic image analysis system in which fat regions were drawn on the image by a food technology expert. Although that technique is cheap and no-contaminant, such manual drawing can be tedious and subject to the expert objectivity. And what is more, although Computer Vision is essentially a non-destroying technique, ham pieces must be destroyed to obtain images with a CCD. So, a fully automatic image analysis would be very desirable and appreciated by the pork industries.

Thresholding techniques are the simplest and most widely used methods to automatically segment images [9]. They are based on the assumption that all pixels whose gray value lies within a certain range belong to one class. Thresholds for that range are automatically established from the image features. Nevertheless, such methods pay no attention to all the spatial information of the image and do not manage acceptably with noisy images, as it has been proved in a previous work [7].

In [8] some thresholding methods was proved to achieve automatic segmentation of ham images in two classes (fat and lean) in order to evaluate their IMF level. Contextual information of the images was incorporated to the proved methods by pre-processing the original images with a rule-based spatial algorithm at different scales and then applying general thresholding techniques to the resultant images. More intuitive pre-processing techniques like median filters and morphological operators were also been tested.

In this paper the same work has made on a new set of images, now with MRI. In the Data set section two different kind of images are described. The correlation between the IMF physical results, chemically obtained, and the Computer Vision data have been calculated for the two sets of images. And in the Results section a comparison between the CCD images and the MRI correlations will be presented. Concluding the practical validation of some of the studied techniques and the rejection of other methodologies, as well as the demonstration that MRI produces better results (better correlations) than CCD images.

2 Data Set

The presented research is based on two different sets of images: images obtained by a CCD camera, and MRI. On the one hand, the CCD database used in the experiment consists in 68 images taken from raw muscle slices of 34 Iberian pig. There are 34 *biceps femoris* and 34 *semimembranosus* images. Slices are digitized at 0.2 mm spatial resolution and 8 bits of depth.

On the other hand, there is a second database of images, formed by MRI Iberian ham images. The images have been acquired using an MRI scanner facilitated by the "Infanta Cristina" Hospital in Badajoz (Spain). The MRI data set is obtained from sequences of T1 images with a FOV (*field-of view*) of 120x85 mm and a slice thickness of 2 mm, i.e. a voxel resolution of 0.23x0.20x2 mm. The total number of MRI used in this work is 36 for the *biceps femoris*, and 36 for the *semimembranosus* muscle. So, the total number of images is 72 for this second database.

Intramuscular fat contents were chemically analyzed for all the slices of the first database, and for the whole muscles in the second set of images. In the experiments, these chemical results were later compared with the Computer Vision outcomes. Considering the two sets of images, the total amount of images is 140.

3 Methods

The behavior of well-known standard thresholding techniques to evaluate fat content in commercial slices of dry-cured Iberian ham has been proved in our previous works [7,8]. Otsu [12], Pun [15], Tsai [21] and multi-level Otsu [16] produced rather low correlation with chemical analysis. But Kapur [11], Johansen & Bille [10] and Kittler [18] methods showed extremely poor correlation.

All the referenced methods above automatically determine a threshold from the image gray level histogram by minimization, maximization or preservation of a criterion function. Hence, Otsu and multi-level Otsu methods maximize criterion functions; Kapur, Johannsen and Bille, Pun and Kittler methods are based on the minimization of functions which depend on a priori and a posteriori image entropy. Tsai method is based on moment's preservation criteria before and after image segmentation.

From all methods, except multi-level Otsu, a single threshold T is determined. It only allows us to classify pixels into two classes. Nevertheless, visual appearance of original images, from the viewpoint of optical density, seems to indicate that pixels could be macroscopically classified into three classes (lean, fat and intermediate, which basically is meat with a high content of water). From this fact, multi-level Otsu method is applied to obtain two thresholds T_1 and T_2 ($T_2>T_1$) and pixels higher than T_2 are classified as fat.

3.1 Practical Application

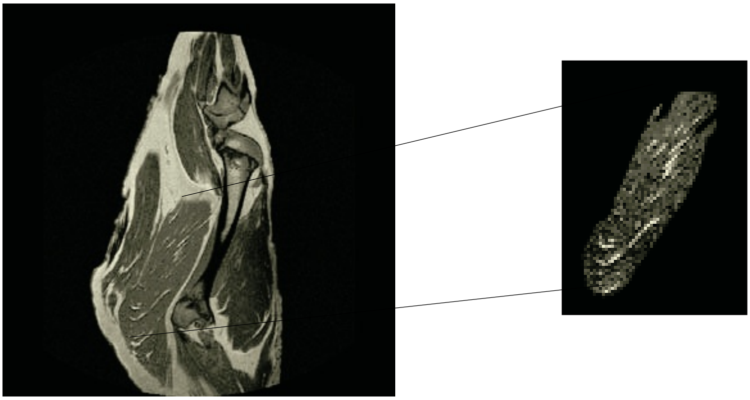
As it has been concluded above, a previous work has proved that multi-level Otsu method is the most suitable methodology to classify this kind of images into three categories. Consequently, this has been chosen as the method to be used in the practical application described in this section.

An experiment has been designed and carried out, in which two different sets of images (CCD and MRI) have been processed with the same methods. CCD images contain spurious noise and fluctuations, which possibly implies a pre-processing stage. On the other hand, this previous stage could have adverse effects for the MRI, which have been obtained avoiding such noises and undesirable effects.

In any case, for both sets of images, different types of preprocessing methods have been used, in an attempt to study their influences on these two heterogeneous groups of images. So, contextual or spatial image information has been incorporated to the developed method pre-processing the original image before applying the multi-level Otsu method to the sample image, *S*.

For the sake of notation, let the original image, *f*, be defined on the discrete set $F=\{x,y/x\in[0,N],y\in[0,M]\}$ where *NxM* is the image size and *f(x,y)* is the gray-level at position *(x,y)*.

Let $S=\{(x,y)/x\in[0,N],y\in[0,M],f(x,y)\neq 0\}$, $S\subset F$, the subset of pixels representing the muscle slice. Then, IMF percentage is obtained dividing the number of fat pixels by the number of pixels in the sample image.



a) Original MRI of a ham

b) Final image, *S*, classified in three regions

Fig. 1. Illustration of Iberian ham MRI.

Three different types of preprocessing approaches have been used: median filters [9], which attenuate spurious noise in the original image, and opening morphological operators [19], for the sake of removing up features in the original image smaller than the kernel size, which could be associated with noise or artifacts. The rule-based algorithm works as follows: let μ_s be the mean gray value on sample image *S*, and μ_{xy} the average gray value of pixels in the image which fall into the kernel. The input image value at each position *(x,y)* remains unchanged when $\mu_{xy>}\mu_s$ and is labeled as background (0 gray level) otherwise in the pre-processed image. The effects of applying this pre-processing are twofold but only with images taken by the CCD. First, spurious noise in sample images is partially removed, and second, pixels that clearly belong to lean are rejected in the subsequent processing.

Finally a fourth experiment has been carried out without any preprocessing stage, to study the advisability of using a preprocessing stage.

For all methods, scale information is added by processing sample images with different kernel sizes. Particularly, two quite different kernel sizes have been considered: 3x3 and 7x7 pixels.

Figure 1 contains an example of MRI. It shows the original MR image (a) and the final classification in three categories (b), using the proposed method (this last one represents the *biceps femoris* muscle). Figure 2 shows the differents experiments carried out with each image. The thick arrow highlights the best obtained performace on the CCD dataset.

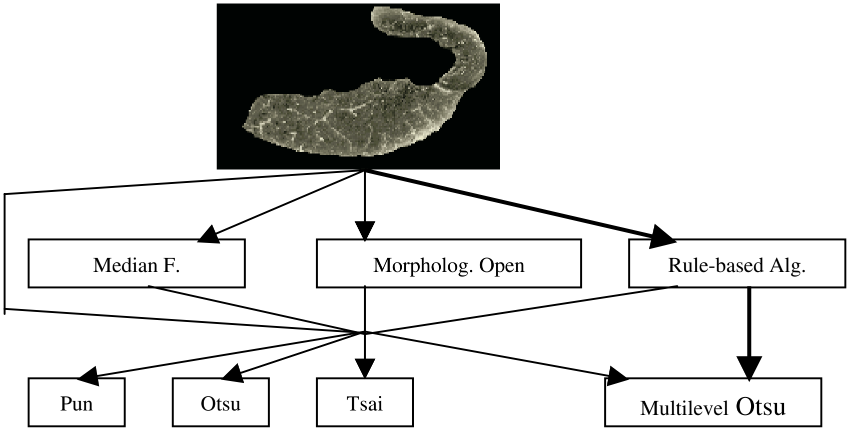


Fig. 2. Schedule of experiments.

4 Results and Discussion

All methods have been proved on the two dataset of images. Tables 1 and 2 show the Pearson correlation of IMF content obtained by the automatic Computer Vision methods and the chemical analysis for *biceps femoris* and *semimembranosus* muscles, respectively. Pearson correlation value mathematically varies between –1 and 1, where a higher value indicates a better estimation of chemical IMF content of the muscles by the computational methods.

Table 1. Pearson coefficient (R) for the Multi-level Otsu method to evaluate IMF content in the *biceps femoris* muscle, for CCD and MRI.

Pre-processing technique	Kernel size	Multi-level Otsu method (CCD)	Multi-level Otsu method (MRI)
-	-	0.23	**0.50**
Median	3	0.27	**0.56**
Filter	7	0.22	0.06
Morhological	3	0.21	-0.53
Opening operator	7	0.10	0.05
Rule-based	3	0.35	-0.58
Algorithm	7	0.32	-0.58

According to the results shown in tables 1 and 2, it is significant that, for the obtained results for the CCD images, the best scores are the values obtained using the rule-based algorithm. These CCD images are very noisy. Water content in the meat generates noise in the image during the acquisition procedure. An intuitive reasoning leads to think that pre-processing techniques, which attenuate spurious noise or remove small features, should improve global performance. In the experiments with the CCD images, only the rule-based algorithm improves performance, especially for the biceps muscle.

Table 2. Pearson coefficient (R) for the Multi-level Otsu method to evaluate IMF content in the *semimembranosus* muscle, for CCD and MRI.

Pre-processing technique	Kernel size	Multi-level Otsu method (CCD)	Multi-level Otsu method (MRI)
-	-	0.00	**0.63**
Median	3	-0.02	**0.60**
Filter	7	-0.06	-0.50
Morhological	3	-0.09	-0.65
Opening operator	7	-0.15	-0.58
Rule-based	3	0.08	-0.66
Algorithm	7	0.11	-0.66

On the other hand, the results obtained for the MRI could be considered as satisfactory, according to the Colton's classification for statistical correlation coefficients [5]. In this case, the best scores are obtained either without any preprocessing stage (0.50 for the *biceps femoris* and 0.63 for the *semimembranosus* muscle), or with the median filter previous stage with 3x3 kernel (0.56 and 0.60 for the *biceps femoris* and *semimembranosus* muscle respectively). The MRI have been acquired using a high quality device, avoiding noise and other undesirable fluctuations. Any of the tested pre-processing methods deteriorate the results, except the median filter pre-processing stage. Therefore, no previous stages are necessary with this type of images. There are significant differences between the two used kernel sizes, being better the 3x3 kernel, maybe due to the high resolution in MRI.

Figure 3 shows the statistical correlation obtained between the four different Computer Vision methods (multi-level Otsu method with the four preprocessing stages) and the chemical data, for the *biceps femoris* and *semimembranosus* muscle. These are the results obtained for MRI. In addition, a tendency line has been added in Figure 3, in order to show the predisposition of the data.

Both chemical and Computer Vision results are certainly comparable for the method that preprocess images using the median filter, and for the method with no preprocessing stage. They appear as fill figures in the graphic, close to the tendency lines. The improvement is especially significant for the *semimembranosus* muscle. Possibly, this could be a consequence of the MRI uniformity.

5 Conclusions

A comparative study among several preprocessing methods in two different sets of images has been presented in this work. Images obtained by a digital camera, generally noisy, have been proved to need a preprocessing stage. On the other hand, the results have revealed that MRI do not require such preprocessing methodologies. The consistent results obtained by the statistical correlation confirm the feasibility of using Computer Vision methods as an alternative to the traditional methods of determining the IMF level of Iberian ham, as the second contribution of the paper.

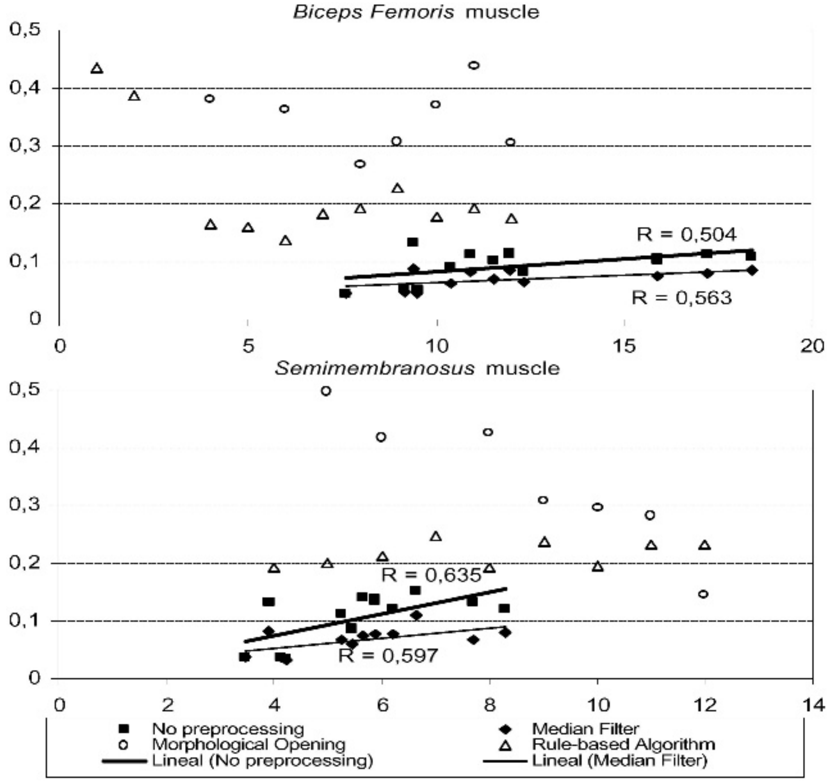


Fig. 3. Pearson correlation coefficients (R) for MRI and 3x3 kernel. X-axis represents IMF level obtained by chemical methods. Y-axis represents the IMF percentage obtained by Computer Vision methods.

Acknowledgments

This research has been funded by the Junta de Extremadura (Regional Government Board) under the IPR98A03P- and 2PR01C025-labeled projects.

References

1. Antequera, T., López-Bote, C.J., Córdoba, J.J., García, C., Asensio, M.A., Ventanas, J. and Díaz, Y., Lipid oxidative changes in the processing of Iberian pig hams, Food Chem., Vol. 45, 105, (1992)
2. Cameron, N.D., Warris, P.D., Forte, J.S. and Enser, M.B., Comparison of Duroc and British Landrace Pigs for Meat and Eating Quality, Meat Science, Vol. 27, 227, (1990)

3. Caro, A., Rodríguez, P.G., Cernadas, E., Antequera, T., Disminución volumétrica del jamón ibérico durante su maduración analizando imágenes de Resonancia Magnética mediante Contornos Activos, Revista Información Tecnológica, Vol. 13-3, 175-180, (2002)
4. Caro, A., Durán, M.L., Rodríguez, P.G., Antequera, T. and Palacios, R., Mathematical Morphology on MRI for the Determination of Iberian Ham Fat Content. Lectures Notes in Computer Science, LNCS-2905, 359-366, (2003)
5. Colton, T.: Statistical in Medicine. Little Brown and Co., Boston – USA, (1974)
6. Davies, A.M.C. and Grant, A. International Journal of Food Science & Technologies, Vol. 22, 191-207, (1987)
7. Durán, M.L., Cernadas, E., Plaza, A., Sánchez, J.M. and Antequera, T., Comparative Study of Segmentation Techniques to Evaluate Fat-Level in Iberian Ham, VIII NSPRIA, Bilbao, Spain. 45-46, (1999)
8. Durán, M.L., Cernadas, E., Plaza, A., Sánchez, J.M., Rodríguez, F., Petrón, M.J., Could Machine Vision Replace Chemical Procedure to Evaluate Fat Content in Iberian Pig Meat? An Experimental Study, 3rd Int. Conf. on Computer Vision, Pattern Recognition, and Image Processing, 256-259, (2000)
9. Haralick and Shapiro, Computer and Robot Vision, Vol. I, Addison-Wesley, (1992)
10. Johannshen, G. and Bille, J., A threshold selection method using information measures, Proceedings, 6th Int. Conference of Pattern Recognition, 140-143, (1982)
11. Kapur, J..N., Sahoo, P.K. and Wong, A.K.C., A New Method for Gray-Level Picture Thresholding Using the Entropy of the Histogram, CVGIP Vol. 29, 273-285, (1985)
12. Otsu, N., A Threshold Selection Method from Gray-Level Histograms, IEEE Trans. Systems, Man, and Cybernetics, Vol. SMC-9 No. 1, (1979)
13. Penfield, M.P., Costello, C.A., McNeil, M.A. and Rienmannn, M.J., Effects of Fat Level and Cooking Methods on Physical and Sensory Characteristics of Reestructure Beef Streaks, Journal Food Qual., Vol 11, 349, (1989)
14. Petrón, M.J., Estudio Comparativo de la Fracción Lipídica de Jamón Fresco en Diferentes Tipos de Cerdo Ibérico, PostGraduate Project, Universidad de Extremadura, (1998)
15. Pun, T. Entropic thresholding: A new approach, CVGIP, Vol. 16, 210-239, (1981)
16. Ritter and Wilson, Handbook of Computer Vision in Image Algebra, CRC Press, (1996)
17. Ruiz J. Estudio de parámetros sensoriales y físico-químicos implicados en la calidad del jamón Ibérico, PhD Dissertation, Universidad de Extremadura, (1996)
18. Sahoo, P.K., Soltani, S. and Wong, A.K.C., A Survey of Thresholding Techniques, CVGIP, Vol. 41, 233-260, (1988)
19. Serra, J., Image Analysis and Mathematical Morphology, Academic Press, (1982)
20. Touraine, B., Vigneror, P., Touraille, C. and Prud'hom, M., Influence des onditions d'elevage sur les characteristiques des carcasses et de la viande d'agneaux Merino d'Arles, Bulletin Technique de l'elevage Ovin, Vol. 4, 29 (1984)
21. Tsai, W-H., Moment-Preserving Thresholding: A New Approach, CVGIP Vol. 29, 377-393, (1985)

Automatic Mask Extraction for PIV-Based Dam-Break Analysis*

Alberto Biancardi[1], Paolo Ghilardi[2], and Matteo Pagliardi[2]

[1] Dept. of Computer and Control Engineering
University of Pavia, Italy
alberto.biancardi@unipv.it
[2] Department of Hydraulics
University of Pavia, Italy
{ghilardip,matteo.pagliardi}@unipv.it

Abstract. The analysis focus on dam breaks stems from their ability to offer a simplified, yet effective workbench for debris flow waves, which in turn are helpful in gaining a deeper understanding of the highly destructive debris flows. High-speed recordings of granular flows arising from a dam-break-like event can be processed to extract useful information about the flow dynamics.

Gradient-based optical-flow techniques cannot compute the correct velocity field as they detect the flow induced by the boundary evolution. Methods that are based on cross-correlation, such as particle imaging velocimetry (PIV), are able to capture the micro-scale flow, but, as they are designed for flows within fixed boundaries, they cannot deal directly with dam-break-caused flows because such flows, by their own nature, exhibit a fast moving boundary.

This paper presents a procedure that is able to compute the evolving background and supply it to a PIV program as a masking region that should be excluded from the computation of the flow velocity field. This improvement leads to reliable results, while reusing existing software.

All the resulting quantities are being used to tune a mathematical model describing the observed flows.

1 Introduction

The need to better understand debris flows comes from the high destructive power that such flows have exhibited in many dreadful occasions: they can exert great impulsive loads on objects they encounter and they are fluid enough to travel long distances or to inundate vast areas. Even commonplace debris flows can denude vegetation, clog drainage ways, damage structures, and endanger humans [1].

In some real world cases debris flows can be triggered by phenomena that are very close to a dam break. Water flows generated by a dam break have

* This work was partially granted by Italian Ministry of Instruction, University and Research under contract PRIN 2003.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 705–712, 2005.

been widely studied and mathematical models for water dam-break waves are available on many textbooks [2]. Compared to water dam-break waves, debris flow waves display a wider variability and, for their mathematical description, they require models with much higher complexity. The set of equations explained in the next section, together with a suitable rheological model, are routinely used by scientists and engineers to describe granular fluid flows.

However, owing to this higher complexity, the values of the parameters used by the models and the structure of models themselves play a critical role in making the simulations exactly describe the real phenomena. This is why laboratory experiments are used to reproduce partially simplified scenarios in order to measure key quantities that can be used as a means both to improve the models and to validate them.

In the specific case of debris flows, as described in the next section, a number of fundamental pieces of information can be gained by determining the velocity vector field. The purpose of this paper is to show how the evolution of such field can be computed from experimental data. The paper will proceed as follows: after presenting the hydrodynamic motivation and our experimental set-up, the outline of the procedure is described; then the determination of the masking region is detailed together with its underlying image-analysis tools. The conclusions and future work will close the paper.

2 Hydrodynamic Motivation

Very often a granular fluid is modelled as a continuum, for which the principle of mass conservation leads to the so-called continuity equation:

$$\frac{d\rho}{dt} + \rho\nabla\cdot \mathbf{v} \tag{1}$$

where ρ is the fluid density (kg/m^3) and $\mathbf{v}$ the velocity vector (m/s).

The density varies with time since the distance between the solids grains varies while the material is flowing. To compute the temporal density variation with equation (1), simultaneous measures of velocities are needed in a wide portion of the fluid to estimate the divergence of the velocity vector v and then the temporal variation of density. The knowledge of density with respect to time is important for an evaluation of the mass rate flow, a quantity often employed in industrial processes.

To test mathematical models of granular fluid flow, further information can be obtained on the same ground. The whole strain rate tensor can be estimated from the same measurements:

$$\dot{\boldsymbol{\varepsilon}} = \frac{1}{2}\left(\nabla\mathbf{v} - \nabla\mathbf{v}^t\right) \tag{2}$$

This tensor is related to the stress tensor which, for a three dimensional flow, is written as follows:

$$\boldsymbol{\sigma} = p\mathbf{I} - 2\mu\left(\dot{\boldsymbol{\varepsilon}} - \frac{1}{3}Tr(\dot{\boldsymbol{\varepsilon}})\mathbf{I}\right) \tag{3}$$

where $\boldsymbol{\sigma}$ is the stress tensor (N/m^2), p is the pressure (N/m^2), $\mathbf{I}$ is the unit tensor, $\dot{\boldsymbol{\varepsilon}}$ is the strain rate tensor (s^{-1}), and μ $(Pa.s)$ is a dynamic viscosity coefficient which, for granular fluids, is a function of the strain rate through a rheological model. The velocity vector $\mathbf{v}$ and the stress tensor $\boldsymbol{\sigma}$ are related also by the momentum equation:

$$\frac{d\mathbf{v}}{dt} + \frac{1}{\rho}\nabla\cdot\boldsymbol{\sigma} = \mathbf{g} \tag{4}$$

where $\mathbf{g}$ is the body force per unit mass vector (m/s^2), which in many cases is equal to the acceleration due to the gravitational field.

2.1 Experimental Set-Up

In the laboratory equipment that has been set up to study dam-break-arisen debris flows, a rectangular flume is used with width $b = 10cm$ and a smooth plastic bed. Debris flows are triggered by the quick opening of a gate, allowing the material accumulated on one side of the gate to free flow to the other side. Several materials are used, from spherical glass beads to almost cylindrical plastic PET grains, with almost constant grain sizes ranging from $1mm$ to $4mm$ and grain densities from $\rho_s = 1285kg/m^3$ (PET) to $\rho_s = 2480kg/m^3$ (glass beads). The gate acceleration was measured from video sequences. Usually the gate is completely removed in a time interval of less than $0.1s$, with an acceleration ranging from 30 to $40m/s^2$.

In our case, each experiment is recorded with high-speed video progressive cameras with digital interface made by Pulnix. These cameras are capable of a sustained rate of 36 million pixels per second that may be arranged into different frame-rates, from 120 to 350 frames per second. Image acquisition by a progressive scan method is particularly suited to the measurements of experimental data because the whole image is captured at once by the camera, while interlaced cameras capture only odd- or even-line sub-frames. The sequences are recorded on a standard PC directly to RAM and then saved to disk in order to keep costs minimal.

An example of the frames recorded during experiments is shown in Fig. 1.

3 Velocity Field Computation

There are a few ways to compute velocity vector fields. They may be grouped into three families: those belonging to the family of gradient-based optical flow; those based on cross-correlation, such as correlation-based optical flow and classical particle image velocimetry (PIV); and those methods that work by tracking single particles or features along their trajectories, such as feature-based optical flow and particle tracking velocimetry (PTV). It is important to note that, while the first two categories supply a regularly-meshed and dense field, tracking solutions give raise to sparse results as they mainly rely on the presence of markers, either particles or features, within the granular flow.

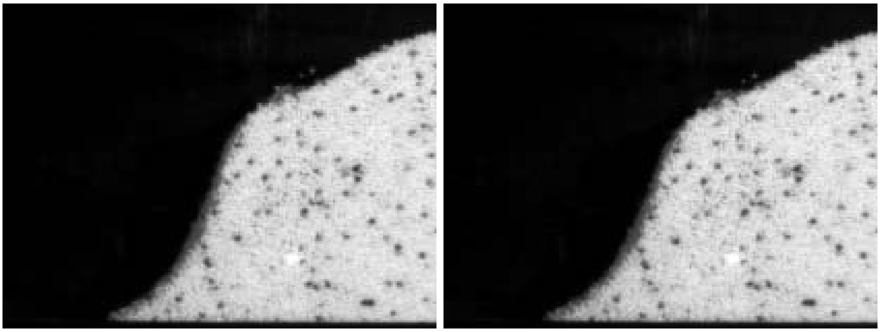


Fig. 1. Two consecutive frames from an experiment recording.

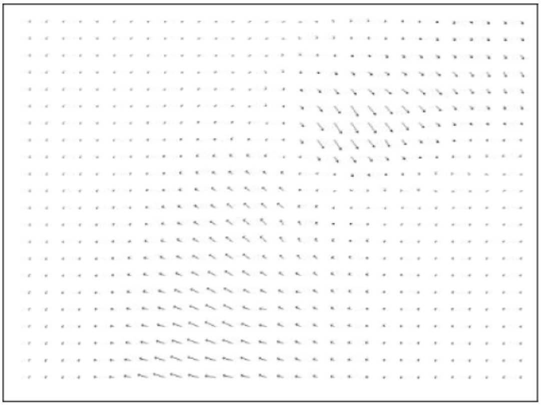


Fig. 2. The optical flow of the example frames

Unfortunately gradient-based optical flow techniques [3] cannot resolve the fine movement of the particles as they are able to see only the macroscopic change caused by the evolution of the particles boundary. The velocity field describes a process where a non-rigid body is pressed on its top and the lower part of its base gets deformed outwardly because of the pressure as shown by Fig. 2.

On the other hand correlation-based PIV methods [4] can get a more precise result because they are designed to catch the movements of many small particles (almost) filling the field of view. The limitation of classical PIV methods, when dealing with flows caused by dam-breaks, is that their implementation is optimised for fluid flows occurring within fixed boundaries while dam-break flows exist because such a fixed boundary is lifted and the once static material is made to move.

MatPIV [5] can compute the velocity field between two consecutive frames and it can be give a mask region that should be excluded from the flow computation. In its current state such a region can only be entered by user interaction – a solution that is adequate as far as such a region does not change for a whole set of experiments, but it is unsatisfactory when the region changes every frame.

The importance of extracting the right masking region is show in Fig. 3, where two velocity fields are compared: the one on the left resulting from the computation without any msking region, the one on the right being computed by taking advantage of the information provided by the masking region.

The problem of extracting the right masking region comes from the presence of dark particles within the white ones. It is necessary to avoid that dark markers

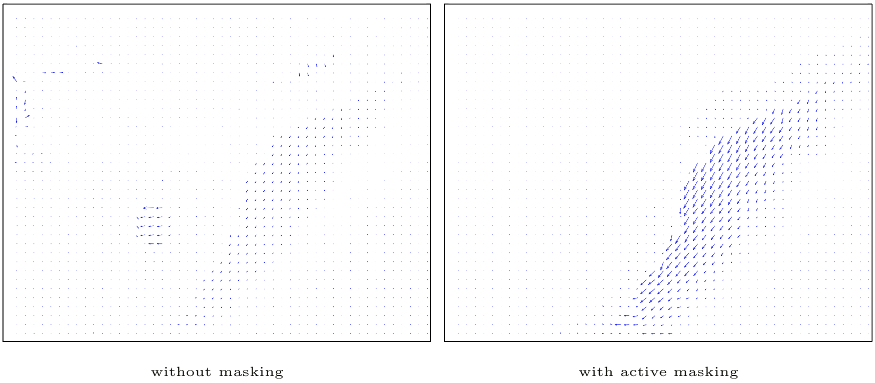


Fig. 3. Comparison between a simulated flow and an experimental flow profile.

near the flow boundary cause an over-extended background determination that prevents the computation of the velocity field inside the flow boundary.

One solution is to use area filters, explained in the following section, to regularize the image and then to binarize it. The resulting region, displayed as white pixels, matches well the background of the example frames as show by Fig. 4.

4 Image Analysis by Connected Components

The procedure implementation to compute the masking region was developed by using data-parallel operators based on connected components. Before proceeding with the description of such procedure, the basic working tools are presented: the notion of connected component, of connected filters, and of data-parallel processing .

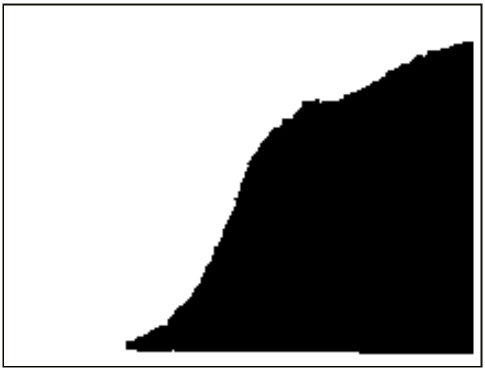


Fig. 4. The output of the background selection on one of the example frames.

4.1 Basic Definitions

The following definitions aim at defining precisely the rather intuitive concept of connected component.

Given an image (of $\mathcal{N} \times \mathcal{M}$ pixels), the concept of adjacent pixels can be formally described using the image planar graph representation [6] where vertices represent pixels and edges exists when and only when two pixels are adjacent.

Domain: An image domain $\mathbf{D}$ is a (rectangular) subset of $\mathbb{N} \times \mathbb{N}$

Discrete image: A discrete image is a mapping of a domain $\mathbf{D}$ into a set of computer representable values I (eg. a subset of $\mathbb{N}, \mathbb{Z}$, or $\mathbb{Q}$):

$$f : \mathbf{D} \to I$$

Relationship: If two pixels $\mathsf{U}, \mathsf{V} \in \mathbf{D}$ are related by the binary relation $\diamond \subseteq \mathbf{D} \times \mathbf{D}$, they will be noted

$$\mathsf{U} \diamond \mathsf{V}$$

Connectivity: For each image pixel a (symmetrical) connectivity relationship $\diamond$ is defined so that

$$\mathsf{U} \diamond \mathsf{V} \textit{ iif } \text{pixels } \mathsf{U} \text{ and } \mathsf{V} \text{ are adjacent}$$

If $\mathsf{V} \in \mathbf{D}$, the (nearest) neighbourhood of V is the subset

$$\mathbf{N} = \{\mathsf{Q} \in \mathbf{D} : \mathsf{V} \diamond \mathsf{Q}\}$$

Connected components: The partitions of $\mathbf{D}$ by $\diamond$ are called connected components. Equivalently, two pixels $\mathsf{U}, \mathsf{V} \in \mathbf{D}$ belong to the same connected component iff there exist a set

$$\mathbf{P} = \{P_0, P_1, \ldots, P_n\}$$

such that

$$P_0 = \mathsf{U}, P_n = \mathsf{V} \text{ and } \forall i \in \mathbb{N}, 0 \leq i < n : P_i \diamond P_{i+1}$$

4.2 Connected Filters

Connected filters are filters designed using connected components. Among connected filters, filters by reconstruction [7] play an important role and collect openings by reconstruction and closings by reconstruction.

In particular we define *opening by reconstruction* any operation that is the composition of any pixel-removing operation composed with a trivial connected opening, which actually reconstructs any connected component that has not been completely removed; on the other hand *closing by reconstruction* is the dual operation in that it is the composition of a pixel-adding operation composed with a trivial connected closing, which completely removes any component which is not entirely preserved. Connected openings and connected closings are also known under the names of geodesic dilations and geodesic erosions [8] or propagations [9] depending on the different points of view they were first introduced.

Filters by reconstruction for grey-scale images are computed by stacking (i.e. adding) the result of their binary counterparts applied to each of the (grey-scale) image cross sections [10]. It is customary to call pores all the small dark subsets surrounded by light areas and to call grains all the small light subsets surrounded by dark areas.

Area filters, which inspired our method, belong to the class of filters by reconstruction. They are chosen among the filters by reconstruction because of their shape-preserving ability while reducing variations among pixel values. Their practical use follows the standard guidelines for alternating filters, thus intermixing opening by reconstruction and closing by reconstruction; in this way they allow the filling of black pores and the removal of white grains. In particular, area openings and area closings use a size criterion for the pixel-removing or pixel-adding operations: any component whose size is less then the required amount is removed

4.3 Data Parallel Programming

Data parallel is a programming paradigm that provides programmers with operators that manipulate images as a whole to perform image transformation. This concept leads to simple code that is close to the application and avoids the explicit manipulation of individual pixels. Even if data parallelism was first used for SIMD architectures, it is now available in some high level languages, such as High Performance Fortran [11], C⋆ [12] or Parallaxis [13], that can be used on workstations or parallel machines of any class.

This approach has proved to be valuable when considering either pointwise operation (thresholding, multi-frame operations, . . .), or local transformations (convolution, morphological operations, . . .).

An extension [14] to this paradigm was proposed so that region-based processing (e.g. moment computation, shape analysis, . . .) could be expressed in terms of data-parallel operations working on connected sets of pixels, which can represent multiple image regions, contours, or other connected parts with respect to the image topology.

The advantages for programmers are reduced development time, more concise code (hence less error prone), higher portability, . . . Additionally by using the connected component extension the processing of multiple objects within a single image is independent of the actual number of objects to be processed, thus simplifying the theoretical development of algorithms as demonstrated by the displacement computation.

The aforementioned area filters, for instance, can be easily implemented in a data-parallel way using very few commands: inside a loop that scans all the grey levels, the sizes of all the connected components is computed and those whose size is less that the user-selected threshold are inverted (grains becoming background and pores filling foreground holes) and the resulting image is accumulated to get the final result at the end of the loop.

5 Conclusions

A procedure to extract the background region of a granular flow was presented. The extracted region is used as a mask thus enabling the computation of the velocity field using standard PIV techniques. The velocity information, measured from high-speed video sequences of granular flow experiments, will be used to tune and validate mathematical and rheological models of debris flows. New simulation programs, based on the smoothed particle hydrodynamics method [15], are being developed and will make use of the computed results.

Acknowledgements

The authors wish to tank Francesca Kovacic who gave a valuable support in running the experiments and analysing the measured data, and Roberto Allieri for his precious assistance in setting up the experimental flumes.

References

1. Takahashi, T.: Debris Flow. Balkema, Rotterdam (1991)
2. Stoker, J.J.: Water Waves. Interscience Publishers, New York (1957)
3. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artificial Intelligence (1981)
4. Raffel, M., Willert, C., Kompenhans, J.: Particle Image Velocimetry. Springer-Verlag (1998)
5. Sveen, J.K.: MatPIV. `http://www.math.uio.no/~ jks/matpiv/` (2004)
6. Serra, J.: Image Analysis and Mathematical Morphology. Academic Press, New York (1982)
7. Salembier, P., Serra, J.: Flat zones filtering, connected operators, and filters by reconstruction. IEEE Transactions on Image Processing **4** (1995) 1153–1160
8. Lantuéjoul, C.: Geodesic segmentation. In Jr., K.P., Uhr, L., eds.: Multicomputers and Image Processing. Academic Press, New York (1982)
9. Duff, M.J.B.: Propagation in cellular logic arrays. In: Proc. Workshop on Picture Data Description and Management. (1980) 259–262
10. Soille, P.: Morphological Image Analysis. Springer-Verlag, Berlin (1999)
11. Forum, H.P.F.: High performance fortran language specification, version 1.0. Technical Report Tech. Report CRPC-TR92225, CRPC-R, Houston (1993)
12. Rose, J.R.: C⋆: a C++-like language for data-parallel computation. Technical Report Tech. Report PL87-8, Thinking Machines Corp., Boston (1987)
13. Bräunl, T.: Parallaxis-III: Architecture independent data parallel processing. IEEE Trans. on Software Engineering (1994)
14. Biancardi, A., Mérigot, A.: Extending the data parallel paradigm with data-dependent operators. Parallel Computing **28** (2002) 995–1021
15. Monaghan, J.J.: Smoothed particle hydrodynamics. Annual Review of Astronomy and Astrophysics **30** (1992) 543–574

Analysis of Meso Textures of Geomaterials Through Haralick Parameters

Margarida Taborda Duarte[1] and Joanne Mae Robison Fernlund[2]

[1] Geomaterisl Research, Calcadinha da Figueira 25-1, 1100-239 Lisbon Portugal
margarida.taborda.duarte@geotextures.com
[2] Department of Land and Water Resources Engineering
Royal Institute of Technology
S-1044 Stockholm Sweden
joanne@kth.se

Abstract. The geomaterials used in this study are granites from Finland with very similar mineral composition. Visual evaluation of the rock texture is done to determine the most significant features of the patterns for the analysis of heterogeneity of meso textures are grain size and grain size spatial distribution. These are compared to results of parameters calculated using image structure analyser. Images are capture with a scanner of the polished slabs that are 9*9 cm in size. The geo textures are expressed by four main parameters: textural entropy, homogeneity, contrast and textural correlation. Reducing the number of parameters to entropy and textural correlation significantly reduce the calculation time. These two parameters are considered to be the most significant. The other two, homogeneity and contrast, can be estimated. The parameter textural correlation yields better results than does textural entropy. Comparison of the analysis of textures visually and using image analysis shows that textural parameters have to be further worked in order to have a better performance.

1 Introduction

This paper presents the first results of the research project "Pattern recognition of geo materials". The goal of the project is to classify the heterogeneity of geomaterials at different scales by quantification of geo textures using image analysis.

For pattern recognition, it is important to deal with parameters that are associated with the most significant features [1] of geomaterials and that these are related to their heterogeneity. Thus, this study is an analysis of geomaterials textures, at a meso scale (from mm to m) using the fourteen Haralick parameters [2]. The goals are: (a) to determine typical values of the textural parameters of an average granite on a meso scale; (b) to explore the physical meaning of these textural parameters, i.e., their capability to indicate the most significant features with respect to the rock heterogeneity.

The analysis of the microstructures using statistical modelling [3] with regard to mechanical behaviour and heterogeneity has been investigated [4], [5], [6]. Some outstanding research dealing with geomaterials are related to the analysis of textures and shape, using image analysis, at different scales. For instance, textural analysis has been applied to estimate rock porosity, within the field of petroleum engineering [7]

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 713–719, 2005.

and the effect of particle shape of coarse aggregates on sieve results, for aggregates industry [8]. Similar to the present study, [7] and [8] have characterized the geomaterials using a more effective method, based on image analysis.

2 Sample Geology

The eight geomaterials, 1) Monola, 2) Balmoral CG (Coarse Grained), 3) Carmen, 4) Eagle, 5) Baltic, 6) Balmoral FG (Fine Grained), 7) Kuru and 8) Porkkala, are from Finland. They are from the geological point of view granites. Their mineral composition shows low variation (Fig. 1). Their average mineral composition is (named Granite in Fig. 1) 61% of feldspar, 31% of quartz, 6% of mica and 2% of other minerals, which is either amphibole or pyroxene.

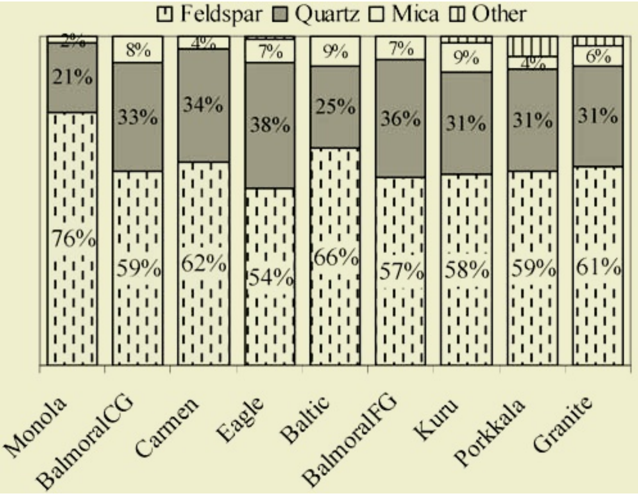


Fig. 1. The mineral composition of the eight geomaterials. Their average composition is shown in the column named Granite.

2.1 Visual Evaluation of the Samples

Visually it is possible to rank the heterogeneity of the rocks and to identify the patterns that are most dissimilar in order to group the most similar ones. Thus, visual evaluation consists in the observation of the most determinant properties of geo patterns to identify what makes them (1) similar to each other (2) more or less heterogeneous.

By analyzing the meso textures (in Fig 2) it is clear that:

a) Kuru has the finniest grain size while Porkkala has the largest grain size.

b) Balmoral has more random spatial distribution of the mineral grains while Carmen has clusters of grains not evenly distributed.

Therefore, the most significant features are the mean grain size and the grain size spatial distribution. These two properties have to be captured by the textural parameters when image analysis is used.

Using visual evaluation to rank the geomaterials, by increasing heterogeneity, the following rank is achieved: Kuru, Balmoral FG, Balmoral CG, Eagle, Carmen, Baltic, Monola, Porkkala.

The results of the evaluation with textural parameters are to be compared with the visual evaluation to conclude about the capacity of the textural parameters to describe the textures.

2.2 Images and Textural Parameters

One image is capture for each one of the eight granites textures. Images are capture by using a scanner and 9*9 cm polished rock slabs. The polish surface is scanned with 300*300 pixels. Image analysis is done using image structure analyser, ISA [9]. For each image the textural entropy, (**TE**), homogeneity (**H**), contrast (**CON**) and textural correlation, (**CORR**) are calculated on images with 256 of grey level. There parameters are calculated with formula **(1), (2), (3), (4).**

$$\text{Textural Entropy, } \mathbf{TE} - \sum_i \sum_j p(i,j)\log(p(i,j)) \tag{1}$$

$$\text{Homogeneity, } \mathbf{H} \sum_i \sum_j \frac{1}{\left|i-j\right|^2} p(i,j), \quad i \neq j \tag{2}$$

$$\text{Contrast, } \mathbf{CON} \sum_i \sum_j (i-j)^2 p(i,j) \tag{3}$$

$$\text{Textural Correlation, } \mathbf{CORR} \frac{\sum_i \sum_j \ (i,j)p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{4}$$

where $p(i,j)$ refers to co-occurrence matrix. This matrix summarizes the textural spatial distribution of intensity of grey levels that occur in images [10].

3 Analysis of the Meso Textures

Results from computation of the meso textures are shown in Table 1. The average of the eight granites is 8.14 of entropy, 0.12 of homogeneity, 460.86 of contrast and 137660.2 of correlation.

Based on further plotting of the data (not presented) it is also verified that there is low variation of textural properties. However larger variations are expected when different geomaterials are analyzed.

To understand the physical meaning, the number of parameters can be reduced to the most significant ones, preferably one or two. This is done by means of the correlation matrix.

Table 1. Results from measurements of textural parameters.

	Entropy	Homogeneity	Contrast	Correlation
Monola	7,62	0,16	71,95	139344,2
BalmoralCG	7,90	0,12	434,62	137091,3
Carmen	7,98	0,13	277,92	115163,5
Eagle	8,05	0,11	423,13	114527,3
Baltic	8,16	0,12	616,95	160350,8
BalmoralFG	8,19	0,11	433,61	110961,9
Kuru	8,58	0,07	956,46	147918,8
Porkala	8,62	0.10	472,26	175923,4
Average granite	8.14	0.12	460.86	137660.2

3.1 Reducing of the Number of Textural Parameters

The correlation matrix gives the linear correlation among pairs of parameters, as shown in Table 2. The correlation is calculated through formulae **(5)**:

$$R = \frac{\frac{1}{n}\left(\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)\right)}{\sigma_x\sigma_Y} \text{ and } -1 \le R \le 1, \tag{5}$$

where n is the number of data; μ_x and μ_Y are the average of x and y data measurements for two textural parameters; σ_x and σ_Y are their variance, respectively. For $|R|=1$, evaluated parameters have a perfect linear correlation and can be modelled with a linear regression. In this study, when ∣R∣ ≥ 65% and statistical significance α<0.1, it is said that there is a clear linear correlation among the two variables. In this case, the two parameters account for similar textural features, and therefore one of them can be discharged.

Table 2. Correlation matrix of the textural parameters with statistical significance (α) correlation in italic.

N=8	Entropy	Homogeneity	Contrast	Correlation
Entropy	1			
Homogeneity	-0.90 α =*0.002*	1		
Contrast	0.77 α =*0.024*	-0.90 α =*0.003*	1	
Correlation	0.49 α =0.213	-0.24 α =0.568	0.34 α =0.144	1

An analysis of Table 2, shows that entropy is strongly correlated to the homogeneity (R=-0.90). Homogeneity, in turn, is strongly correlated to the contrast (R=-0.90). Thus, when entropy is used as a parameter to retrieve information with respect to the geo texture, and the homogeneity and the contrast are measured as well, the latter two parameters are seemingly redundant. In addition, the correlation is not statistically linearly correlated with the other three textural parameters.

The linear models, **H** = -0,069TE + 0,676 and **CON** = -8998H + 1514 are determined by plotting the two correlated parameters (**TE, H**) and (**H, CON**) and fitting a linear model.

Consequently, it was determined that analysis of the meso textures, can be simplified in computation time, by reducing the number of textural parameters to half those initially considered. Entropy and correlation are the two parameters that are used in further analysis while homogeneity and contrast are estimated through the linear models.

3.2 Physical Meaning of Entropy

If we regard the ability of entropy to capture the essence of geo texture for the eight different granite samples it is verified that (Fig. 2):

i. although textural entropy is similar for Kuru and Porkkala, they are disparate in their patterns;
ii. in spite of Monola and Baltic have dissimilar textural entropy, they exhibit analogous patterns;
iii. even though Balmoral FG is similar to the pattern of Balmoral CG, Carmen and Eagle, entropy of the former is not close to the entropy of any of the other rocks types.

Hence, entropy does not account for the overall structure of geo-patterns. This can be explained by the way it is computed. Entropy overtakes texture features at the pixel size and consequently undertakes larger structures. Thus, it is stated, from the above evaluations that textural entropy requires to be further developed to become a better descriptor of geo-pattern.

3.3 Physical Meaning of Correlation

The ultimate goal is to correlate these geo textures with a physical character of the rocks. Several observations can be made (Fig. 3):

i. the geo textures of four of the granites, Balmoral FG, Eagle, Carmen and Balmoral CG, exhibit similar patterns and have similar values of correlation.
ii. Monola and Baltic have more similar values for correlation than of textural entropy.
iii. however, Kuru and Porkkala do not have a sufficiently significant difference in correlation as they appear in pattern, which also has occurred for entropy.

Kuru and Porkkala, are the two that differ most from each other and would therefore be ranked as opposites. Thus, correlation performs better than entropy, but it is not sufficient to distinguish the important features in meso textures.

4 Conclusions

The geo materials used in this study are granites of similar mineral composition from Finland. Both a visual evaluation of their heterogeneity has been made as well as an analysis of their textural parameters using image analysis. The comparison of results

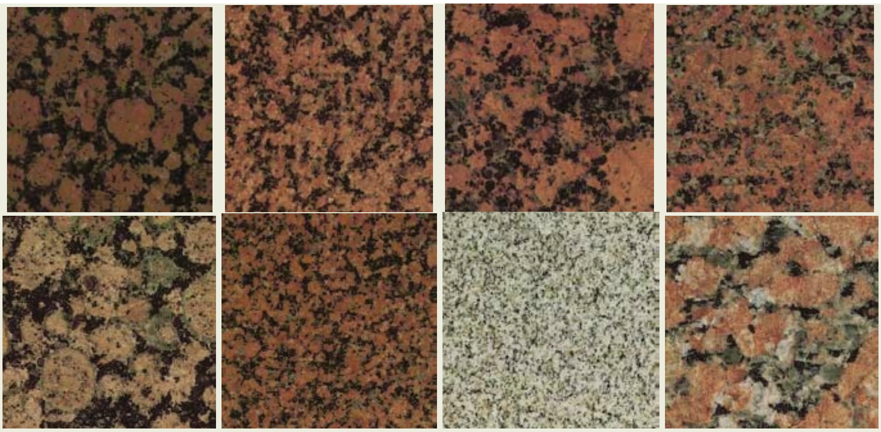


Fig. 2. Ranking of geo textures by increasing entropy, samples 9*9 cm size. **Top from left to right:** Monola TE = 7.62; BalmoralCG TE = 7.90; Carmen TE = 7.98; Eagle TE = 8.05. **Bottom from left to right:** Baltic TE = 8.16; Balmoral FG, TE = 8.19; Kuru, TE = 8.58; Porkkala, TE = 8.62

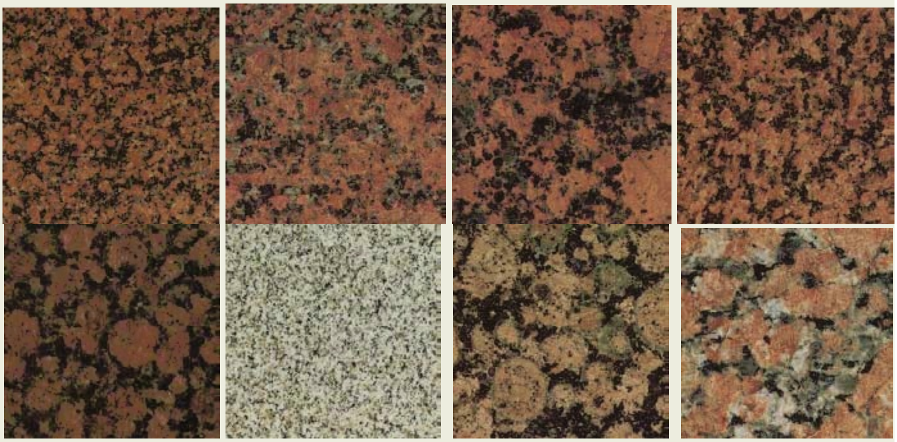


Fig. 3. Ranking of geo textures by increasing correlation, samples 9*9 cm size. **Top from left to write**: Balmoral FG, CORR = 110961,9;Eagle, CORR = 114527,3; Carmen CORR = 115163,5; Balmoral CG, CORR = 137091,3. **Bottom from left to write:** Monola, CORR = 139344,2; Kuru, CORR = 147918,8; Baltic, CORR = 160350,8; Porkkala, CORR = 175923,4

between these two methods revealed substantial differences. The results of this study show that:

I. The analysis of the geo textures, can be simplified by reducing the number of parameters to entropy and correlation, which reduces the compilation time. These are the most significant parameters. The other two parameters, homogeneity and the contrast, can be estimated from linear models, **H** = -0,069**TE** + 0,676 and **CON** = -8998**H** + 1514.

II. the correlation performs better than entropy;

III. the entropy and the correlation have the potential of describing meso textures [11], [12], [13]. Thus, further work has to be done in order to get better performance of the textural parameters .

IV. The grain size and their spatial distribution are identified visually as the remarkable features for the analysis of heterogeneity of meso textures.

References

1. Marques, S.M.: Reconhecimento de padroes, methodos estatisticos e neuronais, (1999) Instituto Superior Tecnico (IST), IST press
2. Haralick, R.M.: Statistical and structural approaches to texture, Proceedings of the IEEE, 67, No. 5, (1979) 786-804
3. Taborda Duarte M., Liu H.Y., Lindqvist P.-A., Kou S.Q., Miskovsky K.: Statistical modelling of the microstructure, accepted in Journal of Materials Engineering and Performance, in press
4. Liu HY, Roquete M, Kou SQ, Lindqvist PA.: Characterization of rock heterogeneity and its numerical verification, Engineering Geology, 72, (2004) 89-119
5. Duarte M. T., Kou S.Q., Lindqvist P.-A.: Miskovsky K., Mechanical heterogeneity of granites based on the weakest link theory, to be submitted
6. Miskovsky K., Taborda Duarte M., Kou S.Q., Lindqvist P.-A.: Influence of the mineralogical composition and textural properties on the quality of course aggregates, Journal of Materials Engineering and Performance, vol. 13, No 2, (2004) 144-150.
7. Lock, P. A., Jing X. D., Zimmerman, R. W., and Schlueter, E. M., 2002, Predicting the permeability of sandstone from image analysis of pore structure, J. Appl. Phys.,10, 6311-6319
8. Fernlund, J.: The effect of particle form on sieve analysis: a test by image analysis. Eng. Geol., 50, (1998) 111–124.
9. Image Structure Analyzer (ISA), Center for Biofilm Engineering's, Montana State University, USA
10. Yang, X., Beyenal, H., Gary, G., Lewandowski, Z.: Quantifying biofilm structure using image analysis, Journal of Microbiological Methods, 39, (2000), 109-119
11. Autio J, Rantanen L., Visa A. and Lukkarinen S.:The classification of rock texture analyses by co-occurrence matrices and the Hough transform, Proc of Geovision, International Symposium on Imaging Applications in Geology, 6th-7th May, 1999 Liege, Belgium
12. Parti M., Cramariuc B., Gabbouj M. and Visa A.: Rock texture retrieval using gray level co-occurrence matrix, Norsig (2002), 5th Nordic Signal Processing Symposium, Tromsø
13. Williams, A.T., Wiltshire and R.J. Thomas, M.C.: Sand grain analysis- Image Processing, textural algorithms and neural nets, Computers & geosciences 24, No 2 (1998), 111-1

Decision Fusion for Target Detection Using Multi-spectral Image Sequences from Moving Cameras

Luis López-Gutiérrez and Leopoldo Altamirano-Robles

National Institute of Astrophysics Optics and Electronics,
Luis Enrique Erro No 1, Santa Maria Tonantzintla, Puebla, 72840 México
luis_david@ccc.inaoep.mx, robles@inaoep.mx

Abstract. In this paper an approach for automatic target detection and tracking, using multisensor image sequences with the presence of camera motion is presented. The approach consists of three parts. The first part uses a motion segmentation method for the detection of targets in the visible images sequence. The second part uses a Gaussian background model for detecting objects presented in the infrared sequence, which is preprocessed to eliminate the camera motion. The third part combines the individual results of the detection systems; it extends the Joint Probabilistic Data Association (JPDA) algorithm to handle an arbitrary number of sensors. Our approach is tested using image sequences with high clutter on dynamic environments. Experimental results show that the system detects 99% of the targets in the scene, and the fusion module removes 90% of the false detections.

1 Introduction

Automatic detection and tracking of objects in image sequences has a lot of applications such as robotic, surveillance and military application, many algorithms have been proposed to solve this problem [1,2]. However, detection and tracking of small, low contrast targets in a highly cluttered environment still remains a very difficult task.

The developed detection and tracking systems uses the probability of detection and the number of false targets to measure the precision; these types of errors can generate false alarm and false rejections. In this single sensor detection system, unfortunately, reducing one type of error comes at the price of increased the other type. We propose an approach to solve the automatic detection problem using more than one sensor (two sensors) and combining the information obtained from these sensors using a decision fusion algorithm. Our principal contribution is improve the target detection and tracking results in image sequences with high clutter, rain, wind and fog without specialization of the algorithms for a specific task; the approach was tested on a set of multi-spectral image sequences with the presence of camera motion.

The paper is organized as follows. Section 2 introduces the models that are considered, and briefly they are described. Section 3 shows an overview of the approach. Sections 4 and 5 describe the algorithms used to detect objects of interest in visible and infrared image sequences respectively. Section 6 describes the method for combining the results obtained by the two algorithms. Several results that validate our approach are reported in section 7, and finally section 8 contains concluding remarks.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 720–727, 2005.

2 Background

Parametric motion model: The parametric motion model w_θ represents the projection of the 3D motion field of the static background [4], where w_θ denotes the modeled velocity vector field and θ the set of model parameters. The parametric motion model is defined at pixel p = (*x,y*) as:

$$\vec{w}_\theta(p) = \begin{pmatrix} a_1 + a_2 x + a_3 y \\ a_4 + a_5 x + a_6 y \end{pmatrix} = \begin{bmatrix} u(p) \\ v(p) \end{bmatrix} \tag{1}$$

Where θ = (a_i), i = 1..6, is the parameter vector to be estimated.

Motion estimation: To estimate a motion model θ_k we use a gradient-based multiresolution robust estimation method described in [4]. To ensure the goal of robustness, we minimize an M-estimator criterion with a hard-redescending function [5]. The constraint is given by the usual assumption of brightness constancy of a projected surface element over its 2D trajectory [6]. The estimated parameter vector is defined as:

$$\hat{\theta} = \underset{\theta}{\text{argmin}}\, \mathrm{E}(\theta) = \underset{\theta}{\text{argmin}} \sum_{p \in R(t)} \rho(DFD_\theta(p)) \tag{2}$$

Where $DFD_\theta(p) = I_{k+1}(p + \vec{w}_\theta(p)) - I_k(p)$, and ρ(*p*) is a function which is bounded for high values of p. The minimization takes advantage of a multiresolution framework and an incremental scheme based on the Gauss-Newton method. More precisely, at each incremental step k (at a given resolution level, or from a resolution level to a finer one), we have: = θ = $\hat{\theta}_k + \Delta\theta_k$. Then, a linearization of $DFD_\theta(p)$ around $\hat{\theta}_k$ is performed, leading to a residual quantity $r_{\Delta\theta k}(p)$ linear with respect to $\Delta\theta_k$:

$$r_{\Delta\theta k(p)} = \nabla I_k(p + \vec{w}_{\hat{\theta} k}(p)) \cdot \vec{w}_{\Delta\theta k}(p) + I_{k+1}(p + \vec{w}_{\hat{\theta} k}(p)) - I_k(p) \tag{3}$$

Where $\nabla I_k(p)$ denotes the spatial gradient of the intensity function at location p and at time k. Finally, we substitute for the minimization of $E(\theta_k)$ the minimization of an approximate expression E_a, which is given by $E_a(\Delta\theta_k) = \sum \rho(r_{\Delta\theta k}(p))$. This error function is minimized using an Iterative-Reweighted-Least-Squares procedure, with 0 as an initial value for $\Delta\theta_k$ [4]. This estimation algorithm allows us to get a robust and accurate estimation of the dominant motion model between two images.

Mixture Gaussian background model: Mixture Models are a type of density model that comprise a number of component functions, usually Gaussian. These component functions are combined to provide a multimodal density [7]. The key idea of background model is to maintain an evolving statistical model of the background, and to provide a mechanism to adapt to changes in the scene. There are two types of background model:

Unimodal model: each pixel is modeled with a single statistical probability distribution (Gaussian distribution) η(X, μt, Σt,), where μt and Σt are the mean value and covariance matrix of the distribution at frame t respectively. Pixels where observed colors are close enough to the background distribution are classified as background points, while those too far away as foreground points.

Multimodal model: a mixture of multiple independent distributions is necessary to model each pixel. Each distribution is assigned a weight representing its priority. A pixel is classified as a background point only if the color observed matches with one of the background distributions. A new distribution of the observation should be imported into the background model if none of the distributions matches it.

Joint Probabilistic Data Association: The Joint Probabilistic Data Association (JPDA) algorithm considers the problem of tracking T targets in clutter [8]. Let $x^t(k)$ ($1\leq t \leq T$) denote the state vectors of each target t at the time of the *k*th measurement. Suppose the target dynamics are determined by known matrices F^t and G^t and random noise vectors $w^t(k)$ as follows:

$$x^t(k+1) = F^t(k)x^t(k) + G^t(k)w^t(k) \tag{4}$$

where t = 1, ... ,T. The noise vectors $w^t(k)$ are stochastically independent Gaussian random variables with zero mean and known covariance matrices. Let m_k denotes the number of validated (or gated) returns at time k. The measurements are determined by

$$z_l(k) = H(k)x^t(k) + v^t(k) \tag{5}$$

where t =1, ... ,T, and *l* =1, ... ,m_k. The H(k) matrix is know, each $v^t(k)$ is a zero-mean Gaussian noise vector uncorrelated with all other noise vectors, and the covariance matrices of the noise vectors vt(k) are know.

The purpose of JPDA is to associate the targets with the measurements and to update those estimates. The actual association of targets being unknown, the conditional estimate is determined by taking a weighted average over all possible associations. An association for the kth observation is a mapping a : {1,...,T}→{0,..., m_k} that associates the target t with the detection a(t), or 0 if no return is associated with it.

Let $\theta_a(k)$ denote the event that “a” is the correct association for the *k*th observation. And $\hat{x}_l^t(k \mid k)$ denote the estimate of $x^t(k)$ given by the Kalman filter on the basis of the previous estimate and the association of the *t*th target with the *l*th return. The conditional estimate $\hat{x}^t(k \mid k)$ for $x^t(k)$ given Z^k is

$$\hat{x}^t(k \mid k) = \sum_{l=0}^{m_k} \beta_l^t(k)\hat{x}_l^t(k \mid k) \tag{6}$$

Where $\beta_l^t(k) = \sum_{a:a(t)=l} P(\Theta_a(k) \mid Z^k)$ is the conditional probability of the event $\theta^t_l(k)$ given Z^k. The set of probabilities $\beta_l^t(k)$ can be computed efficiently as the permanents of a set of sub-matrices.

3 Overview of the Approach

Figure 1 shows an overview of the method. The proposed architecture uses two cameras to get information about scenes with multiple targets and camera motion. The architecture consists of three independent parts.

The first part uses the visible image sequence to calculate the displacement in the whole image, this information is used to compensate the motion originated by the camera, and the targets are detected using spatio-temporal information.

The second part detects the mobile target in the infrared image sequence, in this module the image is preprocessed to eliminate the camera motion found in the previous module, and the targets are detected using a probabilistic Gaussian model. Each part of the algorithm behaves as an expert, indicating possible presence of mobile targets in the scene.

Decision fusion is used to combine the outcomes from all experts, making a final decision.

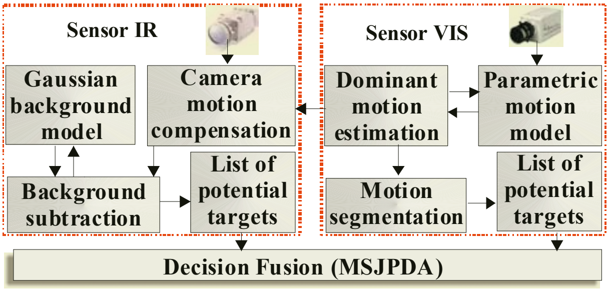

Sensor IR
Gaussian background model
Camera motion compensation
Background subtraction
List of potential targets
Sensor VIS
Dominant motion estimation
Parametric motion model
Motion segmentation
List of potential targets
Decision Fusion (MSJPDA)

Fig. 1. Overview of the approach.

4 Targets Detection in Visible Images

The mobile objects in the visible image sequences are detected performing a thresholding on the motion estimation error, where the mobile objects are the regions whose true motion vector does not conform to the modeled flow vector.

In [10] it is shown through the analysis of the results of different kinds of optical flow estimation algorithms, that $\| \vec{\nabla}\tilde{I}(p) \|^2$ is indeed a proper measure of the reliability of the estimation of the normal flow u_n, thus, the motion error is calculated using the following weighted averaging, which is proposed in [11]

$$Mes_{\hat{\Theta}t}(p) = \frac{\Sigma_{q\in F(p)}\left(\| \vec{\nabla}\tilde{I}(q) \|^2 \times | FD_t(q) |\right)}{Max(\Sigma_{q\in F(p)} \| \vec{\nabla}\tilde{I}(q) \|^2, n \times G_m^2)} \tag{7}$$

Where F(p) is a small neighborhood around p which contains n points, and G_m is a constant which accounts for noise in the uniform areas. An interesting property of this local measure is the following: let us suppose that the pixel p and its neighborhood undergoes the same displacement of magnitude □ and direction $\vec{u}$. In [3] there were derived two bounds l(p) and L(p) such that, whatever the direction $\vec{u}$ might be, the following inequality holds:

$$0 \leq l(\mathrm{p}) \leq Mes_{\hat{\Theta}t}(p) \leq L(\mathrm{p}) \tag{8}$$

The bounds used in the experiments are given by:

$$\begin{cases} l(p) = \eta\delta\sqrt{\lambda'_{\min}(1-\lambda'_{\min})} \\ L(p) = \delta\sqrt{1-\lambda'_{\min}} \end{cases} \text{ with } \eta = \frac{\Sigma_{q\in F(p)} \| \vec{\nabla}\tilde{I}(q) \|^2}{Max(\Sigma_{q\in F(p)} \| \vec{\nabla}\tilde{I}(q) \|^2, n \times G_m^2)} \text{ and } \lambda'_{\min} = \frac{\lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}$$

Where $\lambda_{\min}$ and $\lambda_{\max}$ are respectively the smallest and highest eigenvalues of the following matrix (with $\nabla\tilde{I}(q) = (\tilde{I}_x(q), \tilde{I}_y(q))$:

$$M = \begin{pmatrix} \Sigma_{q\in F(p)}\tilde{I}_x(q)^2 & \Sigma_{q\in F(p)}\tilde{I}_x(q)\tilde{I}_y(q) \\ \Sigma_{q\in F(p)}\tilde{I}_x(q)\tilde{I}_y(q) & \Sigma_{q\in F(p)}\tilde{I}_y(q)^2 \end{pmatrix} \tag{9}$$

5 Targets Detection in Infrared Images

The mobile objects in the infrared sequence are probabilistically determined using a Gaussian model of the background [7], so that everything that does not normally occur in the background is viewed as a potential target. The proposed target detection algorithm has four basic steps:

Motion Compensation The image sequence is preprocessed to eliminate the camera motion, this preprocessing step use the information about the dominant motion calculated in the last module.

Model Generation The homogeneous background is described using Gaussian models, where each pixel is modeled like a mixture of 3 distributions.

$$x \approx \sum_{i=1}^{3} \pi_i N(\mu_i, \sigma_i^2) \tag{10}$$

Model Optimization The model and model size are optimized using training data.

Target Detection In this step the distributions are ordered using the value of the weight assigned to each distribution and the covariance of the distribution. If the weight of the distribution is bigger than a predefined threshold them the distribution is considered background. If the pixel value is accurately described by the model them the pixel is considered an element of the background.

6 Decision Fusion

The first and second parts of the approach each behave as experts indicating the possible position of mobile targets in the scene. The final decision is reached by fusing the results of these experts.

Figure 2 shows the sequential Multi-Sensor Data Fusion architecture [12] used to combine the individual target detecting results. The initial state of the tracking algorithms is obtained using a weighted "k out of N" voting rule. The combination of the measurements is done; making Ns (Number of sensors in the system) repetitions of the JPDA algorithm (see section 2).

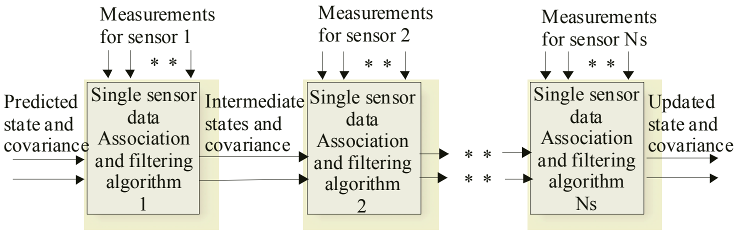

Measurements for sensor 1
* *
Measurements for sensor 2
* *
Measurements for sensor Ns
* *
Predicted state and covariance
Single sensor data Association and filtering algorithm 1
Intermediate states and covariance
Single sensor data Association and filtering algorithm 2
* *
* *
Single sensor data Association and filtering algorithm Ns
Updated state and covariance

Fig. 2. Multi-sensor Fusion Architecture.

The fusion algorithm works on the basis of the following equations.

Let m_{ki}, i = 1, 2, . . . ,Ns, be the number of validated reports from each sensor i at time k. The measurements are determined by

$$z_l^i(k) = H_i(k)x^t(k) + v_i^t(k) \tag{11}$$

where t = 1, . . . ,T, i =1, . . . , Ns, and *l* =1, . . . m_{ki}. The measurement $z_l^i(k)$ is interpreted as the *l*th measurement from the *i*th sensor at time k. Generalizing from the single-sensor case, the $H_i(k)$ matrices are known, and $v_i^t(k)$ are stochastically independent zero-mean Gaussian noise vectors with known covariance matrices. The observation at time k is now

$$Z(k) = (z_1^1(k),\ldots,z_{mk1}^1(k), z_1^2(k),\ldots z_{mk1}^2(k),\ldots,z_1^{Ns}(k),\ldots,z_{mk1}^{Ns}(k)) \tag{12}$$

The conditional estimate of the fusion algorithm is given by:

$$\hat{x}^t(k \mid k) = \sum_L \beta_L^t(k)\hat{x}_L^t = \sum_L \prod_{i=1}^{Ns} \beta_L^t(k)\hat{x}_L^t \tag{13}$$

Where the sums are over all possible sets of associations L with target t.

7 Results

In this section, we will show the experimental results of our approach. The algorithm is tested with a database of 3 multispectral image sequences. The weather conditions were winds of 30 to 70 km/hour, and variable lighting.

Table 1 shows the principal features of each sequence and the results of the two first blocks, in the table Pd is the probability of detection and NFt is the average number of false targets per image. The figures 3(a), 4(a) and 5(a) show images of the Boat, People and Pruning machine sequence respectively. In the first experiment the camera was mount in a maritime platform, in this sequence are presents the motion of the camera, the targets motion and the motion of the sea. In the second and third experiment the camera stayed static, these sequences contain a great amount of noise and low contrast. By applying the algorithms described in section 4 and 5 the targets are detected in each frame, figures 3(b) and (c) show the target detection results using the Boat sequence, figures 4(b) and (c) show the results using the People sequence, whereas in figure 5(b) and (c) are shown the results in the Pruning machine sequence.

In these figures the false detections are showed with a green circle, whereas the detections obtained from the motion segmentation algorithm are marked with a red circle; the detections from the probabilistic model are marked with a white rectangle and the detections obtained from the fusion method are showed whit a circle. These results show that the system can detect any type of mobile targets in the presence of adverse weather condition.

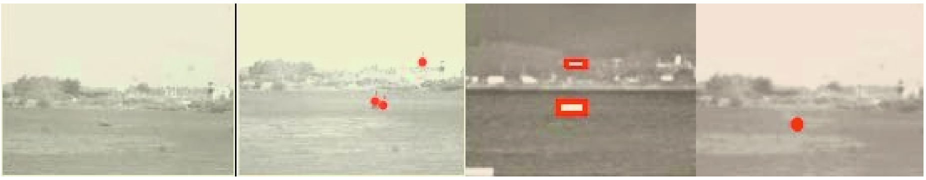
In table 2, results after the decision fusion are shown. In these sequences, the fusion improves results. The data association step in the fusion module reduces the number of false targets creating gating regions and considering just the measurements that fall in that region. The fusion module improves the target state estimation by processing sequentially the sensors detection, in this module if the target was not detected by one sensor, the information about it stays and the following sensor is processing, this way to combine the information improves the probability of detection, because the target must be loosed in all sensors to lose it in the fusion decision result. Figure 3(d), 4(d) and 5(d) shows these results graphically.

Table 1. Results of different experts.

equenc	Size	Frame	Targets	Sensor	Time rocessin	Pd (%)	NFt
Boat	640x480	200	1	Visible	1.8 seg	98	6.2
				Infrared	0.04 seg	97	0.9
People	640x480	200	3	Visible	1.9 seg	95	5.8
				Infrared	0.03 seg	96	0.4
Pruning machine	640x480	200	1	Visible	1.7 seg	97	5.1
				Infrared	0.04 seg	96	0.3

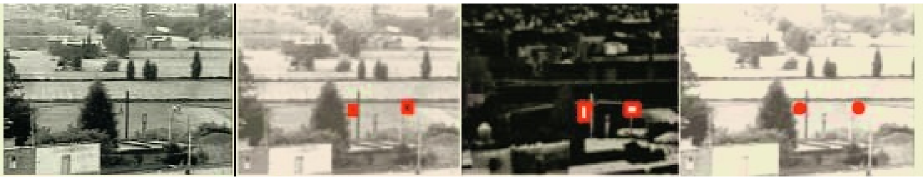
Table 2. Results after fusion.

Sequence	Processing average time	Pd (%)	NFt
Boat	2.0 seg.	100	0.5
People	2.15 seg.	99	0.1
Pruning machine	1.98 seg.	99	0.2



(a) Sequence at k=50. (b) Motion segmentation result. (c) Background model result. (d) Detection after Fusion

Fig. 3. Target detection results in the boat sequence.

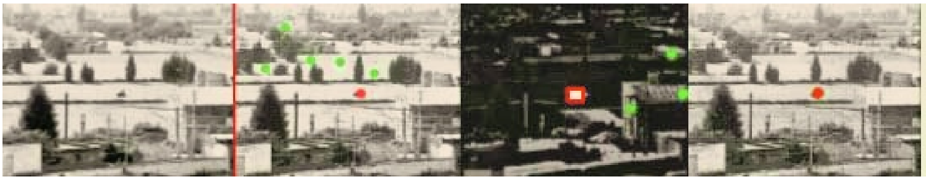


(a) Sequence at k = 50. (b) Motion segmentation result. (c) Background model result. (d) Detection after fusion

Fig. 4. Target detection results in the people sequence.

8 Conclusions

In this paper an approach to improve target detection process using decision fusion is proposed. The approach was tested using multi-spectral image sequences from moving cameras. Experimental results show that targets detection algorithms detects in average 97% of the targets in the worse case, and in the better one detects 99.5%. The fusion module detects in the worst case 99% of the targets and 100% in the better one, while the 90% of the false targets are removed. This results show the advantages of this approach for automatic detection and tracking. It has been shown that this approach performs better that either tracker in isolation. Most importantly the tracking performance is improved without specialization of the tracking algorithms for a specific task; it remains to develop an algorithm to handle target occlusion and to reduce the processing time.



(a) Sequence at k = 50. (b) Motion segmentation result. (c) Background model result. (d) Detection after fusion

Fig. 5. Target detection results in the Pruning machine sequence.

References

1. Wang, D.; Unsupervised Video Segmentation Based on Water-sheds and Temporal Tracking; Trans. Circuits Syst Video Technology, vol 8, 1998, pp 539-546.
2. Foresti, G.L.; Object Recognition and Tracking for Remote Video Surveillance; Trans. Circuits Syst. Video Technol., vol 9, 1999, pp 1045-1062.
3. J. Odobez, P. Bouthemy. Direct incremental model-based image motion segmentation analysis for video analysis Signal Processing. Vol 66, pp 143-155, 1998.
4. J. Odobez, P. Bouthemy. Robust Multiresolution Estimation of Parametric Motion Models. JVCIR, 6(4) pp 348-365, 1995.
5. P.J. Hubert. Robust statistics. Wiley, 1981.
6. Horn, Shunck. Determining optical flow. Artificial Intelligence, vol 17 pp 185-203, 1981
7. C. Stauffer, Adaptive background mixture models for real-time tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 246-252, 1999.
8. Bar-Shalom, T. Fortmann. Tracking and Data Association, Academic P., San Diego, 1988.
9. E. Waltz and J. Llinas, Handbook of Multisensor data fusion, CRC Press, 2001.
10. J. Barron, D Fleet, S. Bauchemin. Performance of optical flow techniques. International Journal of Computer Vision. 12(1) pp 43-77, 1994.
11. M. Irani, B. Rousso, S. Peleg. Computing occluding and transparent motion. Intern. J.Comput. Vis. 12(1) pp 5-16, 1994.
12. L. Pao, S. O'Neil. Multisensor Fusion Algorithms for Tracking. Proc. of American Control Conference. pp. 859--863, 1993.

Author Index