

Abderrahim Elmoataz
Olivier Lezoray
Fathallah Nouboud
Driss Mammass (Eds.)

LNCS 5099

Image and Signal Processing

3rd International Conference, ICISP 2008
Cherbourg-Octeville, France, July 2008
Proceedings



Springer

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Alfred Kobsa

University of California, Irvine, CA, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

University of Dortmund, Germany

Madhu Sudan

Massachusetts Institute of Technology, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max-Planck Institute of Computer Science, Saarbruecken, Germany

Abderrahim Elmoataz Olivier Lezoray
Fathallah Nouboud Driss Mammass (Eds.)

Image and Signal Processing

3rd International Conference, ICISP 2008
Cherbourg-Octeville, France, July 1 - 3, 2008
Proceedings



Springer

Volume Editors

Abderrahim Elmoataz

Olivier Lezoray

Université de Caen Basse-Normandie

GREYC UMR CNRS 6072 ENSICAEN

6 Bd. Maréchal Juin

14050 Caen, France

E-mail: {abderrahim.elmoataz-billah, olivier.lezoray}@unicaen.fr

Fathallah Nouboud

Université de Québec à Trois-Rivières

C.P. 500 Trois-Rivières G9A 5H7, Québec, Canada

E-mail: fathallah.nouboud@uqtr.ca

Driss Mammass

Université IbnZohr

BP.8106 Hay Dakhla, Agadir Morocco

E-mail: driss_mammass@yahoo.fr

Library of Congress Control Number: 2008929562

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

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

ISSN 0302-9743

ISBN-10 3-540-69904-X Springer Berlin Heidelberg New York

ISBN-13 978-3-540-69904-0 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2008, corrected publication 2021

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12319916 06/3180 5 4 3 2 1 0

Preface

ICISP 2008, the International Conference on Image and Signal Processing, was the third ICISP conference, and was held in Cherbourg-Octeville, Normandy, France. Historically, ICISP is a conference resulting from the actions of researchers of Canada, France and Morocco. Second and first editions of ICISP were held in Agadir, Morocco in 2003 and 2001. ICISP 2008 was sponsored by EURASIP (European Association for Image and Signal Processing) and IAPR (International Association for Pattern Recognition).

The response to the call for papers for ICISP 2008 was encouraging. From 193 full papers submitted, 70 were finally accepted (48 oral presentations, and 22 posters). The review process was carried out by the Program Committee members; all are experts in various image and signal processing areas. Each paper was reviewed by at least two reviewers, and also checked by the conference Co-chairs. The quality of the papers in these proceedings is attributed first to the authors, and second to the quality of the reviews provided by the experts. We would like to thank the authors for responding to our call, and we thank the reviewers for their excellent work.

We were very pleased to be able to include in the conference program keynote talks by four world-renowned experts: Joachim Weickert, Full Professor of Mathematics and Computer Science at Saarland University, Germany, where he heads the Mathematical Image Analysis Group; Godfried T. Toussaint, Professor Emeritus at the School of Computer Science and the Centre for Interdisciplinary Research in Music Media and Technology at McGill University, Montreal, Quebec, Canada; Driss Aboutajdine, Full Professor at the Science Faculty of Rabat University, Morocco; and David Tschumperlé, permanent CNRS researcher in the Image group of the GREYC research laboratory, Caen, France.

We would like to thank Amal Mahboubi, the web master of the conference for maintaining the Web pages and Agnès Zannier for her administrative assistance. We would also like to thank the members of the Local Committee for their advice and help. We are also grateful to Springer's editorial staff for supporting this publication in the LNCS series.

Finally, we were very pleased to welcome all the participants to this conference. For those who did not attend, we hope this publication provides a good view into the research presented at the conference, and we look forward to meeting you at the next ICISP conference.

April 2008

Abderrahim Elmoataz
Olivier Lezoray
Fathallah Nouboud
Driss Mammass

ICISP 2008 – International Conference on Image and Signal Processing

General Chair

Abderrahim Elmoataz
Université de Caen Basse-Normandie, France
abderrahim.elmoataz-billah@unicaen.fr

General Co-chair

Fathallah Nouboud
Université du Québec à Trois-Rivières, Québec, Canada
Fathallah.Nouboud@uqtr.ca

Program Committee Chair

Olivier Lezoray
Université de Caen Basse-Normandie, France
olivier.lezoray@unicaen.fr

Program Committee Co-chair

Driss Mammass
Université Ibn Zohr, Morocco
driss.mammass@yahoo.fr

Webmaster

Amal Mahboubi
Université de Caen Basse-Normandie, France
amal.mahboubi@unicaen.fr

Conference Secretary

Agnès Zannier GREYC CNRS UMR 6072
agnes.zannier@greyc.ensicaen.fr

Sponsoring Institutions

Institut Universitaire de Technologie Cherbourg Manche, France
Université de Caen Basse-Normandie, France
ENSICAEN, France
Centre National de la Recherche Scientifique, France
Centre National de la Recherche Scientifique, Délégation Normandie, France
Conseil Régional de Basse-Normandie, France
Conseil Général de la Manche, France
Communauté Urbaine de Cherbourg, France
Pôle Transactions Électroniques Sécurisées, France

Local Organizing Committee

S. Bougleux	Ceremade, Université Paris Dauphine, France
L. Brun	Ensicaen, France
Y. Chahir	Université de Caen Basse-Normandie, France
C. Charrier	Université de Caen Basse-Normandie, France
J. Fadili	Ensicaen, France
Z. Lakhdari	Université de Caen Basse-Normandie, France
A. Mahboubi	Université de Caen Basse-Normandie, France
P. Makany	Université de Caen Basse-Normandie, France
S. Schpp	Université de Caen Basse-Normandie, France
V.-T. Ta	Université de Caen Basse-Normandie, France

Program Committee

T. Adali	University of Maryland Baltimore County, USA
J.M. Aroza	Universidad de Granada, Spain
A. Belaïd	Université Nancy 2, France
M. Bellafkikh	INPT - Rabat, Morocco
B. Boufama	Université de Windsor, Canada
A. Chalifour	Université du Québec à Trois-Rivières, Canada
W.-K. Cham	Chinese University of Hong Kong, China
L. Chen	Ecole Centrale de Lyon, France
D. Chiadmi	EMI-Rabat, Morocco
L. Cohen	Ceremade, Université Paris Dauphine, France
B. Coll	Universitat de les Illes Balears, Spain
J. Crespo	Universidad Politécnica de Madrid, Spain
F. Deschenes	Université du Québec en Outaouais, Canada
T. Deserno	Aachen University of Technology, Germany
F. Escolano	University of Alicante, Spain
A. Evans	University of Bath, UK
D. Gorodnichy	National Research Center, Canada
M. Gurcan	Ohio State University, USA

R. Harba	Ecole Polytechnique d'Orléans, France
S. Jehan-Bresson	ENSICAEN, France
T. Jiang	The Chinese Academy of Sciences, China
Z. Kato	University of Szeged, Hungary
M.L. Kherfi	Université du Québec à Trois-Rivières, Canada
D.-S. Kim	Hanyang University, Seoul, Korea
W. Kropatsch	Vienna University of Technology, Austria
D. Laurendeau	Université Laval, Québec, Canada
D. Lévesque	Université de Sherbrooke, Canada
R. Malgouyres	Université Clermont-Ferrand, France
M. Melkemi	Université Haute Alsace, France
S. Metari	Université de Sherbrooke, Canada
A. Mitiche	Inrs-Energie, Canada
M. Nikolova	CMLA ENS Cachan, France
G. Peyré	Ceremade, Université Paris Dauphine, France
I. Pitas	University of Thessaloniki, Greece
F. Preteux	INT - Evry, France
O. Roudies	EMI - Rabat, Morocco
E. Rubio Royo	Universidad Las Palmas Gran Canaria, Spain
H. Sahli	Vrije Universiteit Brussel, Belgium
A. Sbihi	Université Ibn Tofail - Kénitra, Morocco
B. Smolka	Silesian University of Technology, Poland
J.-L. Starck	CEA, France
A. Torsello	University of Venice, Italy
E. Uchino	Yamaguchi University, Japon
M. Vento	University of Salermo, Italy
L. Wang	University of Melbourne, Australia
D. Zhang	Honk Kong Polytechnic University ,China
D. Ziou	Université Sherbrooke, Canada

Table of Contents

Image Filtering

Discrete Pulse Transform of Images	1
<i>Roumen Anguelov and Inger Fabris-Rotelli</i>	
Constrained Unsharp Masking for Image Enhancement	10
<i>Radu Ciprian Bilcu and Markku Vehvilainen</i>	
Discrete Wavelet Diffusion for Image Denoising	20
<i>Kashif Rajpoot, Nasir Rajpoot, and J. Alison Noble</i>	
Dynamic Color Texture Modeling and Color Video Decomposition Using Bounded Variation and Oscillatory Functions	29
<i>Mathieu Lugiez, Michel Ménard, and Abdallah El-Hamidi</i>	
RETRACTED CHAPTER: Reinstating Floyd-Steinberg: Improved Metrics for Quality Assessment of Error Diffusion Algorithms	38
<i>Sam Hocevar and Gary Niger</i>	
A Neuro Fuzzy Model for Image Compression in Wavelet Domain	46
<i>Vipula Singh, Navin Rajpal, and K. Srikanta Murthy</i>	

Image Segmentation

A People Counting System Based on Dense and Close Stereovision	59
<i>Tarek Yahiaoui, Cyril Meurie, Louahdi Khoudour, and François Cabestaing</i>	
Topology-Preserving Discrete Deformable Model: Application to Multi-segmentation of Brain MRI	67
<i>Sanae Miri, Nicolas Passat, and Jean-Paul Armspach</i>	
A Real Time Human Detection System Based on Far Infrared Vision	76
<i>Yannick Benerezeth, Bruno Emile, Hélène Laurent, and Christophe Rosenberger</i>	

Computer Vision

Uniqueness Filtering for Local Feature Descriptors in Urban Building Recognition	85
<i>Giang Phuong Nguyen and Hans Jørgen Andersen</i>	
Dynamic Human Fatigue Detection Using Feature-Level Fusion	94
<i>Xiao Fan, Bao-Cai Yin, and Yan-Feng Sun</i>	

Real Time Monocular Depth from Defocus	103
<i>Jean-Vincent Leroy, Thierry Simon, and François Deschenes</i>	
Dynamic Skin Detection in Color Images for Sign Language Recognition	112
<i>Michał Kawulok</i>	
Depth Estimation by Finding Best Focused Points Using Line Fitting	120
<i>Aamir Saeed Malik and Tae-Sun Choi</i>	
Improved Phase Correlation Matching	128
<i>Javed Ahmed and M. Noman Jafri</i>	
A New Approach to Estimate Fractal Dimension of Texture Images	136
<i>André R. Backes and Odemir M. Bruno</i>	

Biomedical Applications

Digital Signal Processing Techniques for Gene Finding in Eukaryotes	144
<i>Mahmood Akhtar, Eliathamby Ambikairajah, and Julien Epps</i>	
A CAD System for Long-Bone Segmentation and Fracture Detection	153
<i>Martin Donnelley, Greg Knowles, and Trevor Hearn</i>	
Comparison of Image Restoration Methods for Bioluminescence Imaging	163
<i>Smail Akkoul, Roger Ledee, Remy Leconge, Christophe Leger, Rachid Harba, Sabrina Pesnel, Stéphanie Lerondel, Alain Lepape, and Luis Vilcahuaman</i>	
Texture-Based Multiscale Segmentation: Application to Stromal Compartment Characterization on Ovarian Carcinoma Virtual Slides	173
<i>Nicolas Signolle, Benoît Plancoulaine, Paulette Herlin, and Marinette Revenu</i>	
Combination of Image Registration Algorithms for Patient Alignement in Proton Beam Therapy	183
<i>Rachid Belaroussi and Guillaume Morel</i>	
Speech Signal Processing Based on Wavelets and SVM for Vocal Tract Pathology Detection	192
<i>P. Kukharchik, I. Kheidorov, E. Bovbel, and D. Ladeev</i>	

Feature Extraction

Pair Correlation Integral for Fractal Characterization of Three-Dimensional Histograms from Color Images	200
<i>Julien Chauveau, David Rousseau, and François Chapeau-Blondeau</i>	

Accelerating Integral Histograms Using an Adaptive Approach..... <i>Thomas Müller, Claus Lenz, Simon Barner, and Alois Knoll</i>	209
Volume Estimation of Small Particles Using Three-Dimensional Reconstruction from Multiple Views <i>Eduardo Castillo-Castaneda and Christelle Turchiuli</i>	218
A Multivariate Hit-or-Miss Transform for Conjoint Spatial and Spectral Template Matching <i>Jonathan Weber and Sébastien Lefèvre</i>	226
Blur Insensitive Texture Classification Using Local Phase Quantization <i>Ville Ojansivu and Janne Heikkilä</i>	236
An Algorithm to Determine the Chromaticity Under Non-uniform Illuminant <i>Sivalogeswaran Ratnasingam and Steve Collins</i>	244
Transients Detection in the Time-Scale Domain <i>V. Bruni and D. Vitulano</i>	254
SIFT Based Ball Recognition in Soccer Images <i>Marco Leo, Tiziana D’Orazio, Paolo Spagnolo, Pier Luigi Mazzeo, and Arcangelo Distante</i>	263
Evaluation Protocol for Localization Metrics: Application to a Comparative Study <i>Baptiste Hemery, Hélène Laurent, Christophe Rosenberger, and Bruno Emile</i>	273
Pattern Recognition	
Classification of Proteomic Signals by Block Kriging Error Matching ... <i>Tuan D. Pham, Dominik Beck, Miriam Brandl, and Xiaobo Zhou</i>	281
Grip-Pattern Recognition in Smart Gun Based on Likelihood-Ratio Classifier and Support Vector Machine <i>Xiaoxin Shang and Raymond N.J. Veldhuis</i>	289
Chirplet Transform Applied to Simulated and Real Blue Whale (<i>Balaenoptera Musculus</i>) Calls <i>Mohammed Bahoura and Yvan Simard</i>	296
Illumination Invariant Face Recognition Under Various Facial Expressions and Occlusions <i>Tiwuya H. Faaya and Önsen Toygar</i>	304

A Version of the FastICA Algorithm Based on the Secant Method Combined with Simple Iterations Method	312
--	-----

Doru Constantin and Luminita State

SVM-Based Face Recognition Using Genetic Search for Frequency-Feature Subset Selection	321
--	-----

Aouatif Amine, Mohammed Rziza, and Driss Aboutajdine

Graph-Based Representations

A Path Following Algorithm for Graph Matching	329
---	-----

Mikhail Zaslavskiy, Francis Bach, and Jean-Philippe Vert

Statistically Valid Graph Representations of Scale-Space Geometry	338
---	-----

Tomoya Sakai and Atsushi Imiya

Hierarchical Matching Using Combinatorial Pyramid Framework	346
---	-----

Luc Brun and Jean-Hugues Pruvot

Motion Detection and Estimation

Multi-cue Facial Feature Detection and Tracking	356
---	-----

Jingying Chen and Bernard Tiddeman

Body Language Based Individual Identification in Video Using Gait and Actions	368
---	-----

Y. Pratheepan, P.H.S Torr, J.V. Condell, and G. Prasad

Human Hand Gesture Recognition Using Motion Orientation Histogram for Interaction of Handicapped Persons with Computer	378
--	-----

Maryam Vafadar and Alireza Behrad

3D Human Motion Reconstruction Using Video Processing	386
---	-----

Nadiya Roodsarabi and Alireza Behrad

New Interfaces

Personal Recognition Using Single-Sensor Multimodal Hand Biometrics	396
---	-----

Andreas Uhl and Peter Wild

Face Tracking for Spatially Aware Mobile User Interfaces	405
--	-----

Jari Hannuksela, Pekka Sangi, Markus Turtinen, and Janne Heikkilä

Markerless Augmented Reality Using Image Mosaics.....	413
---	-----

Pietro Azzari, Luigi Di Stefano, Federico Tombari, and Stefano Mattoccia

Retracted: A Novel Text-Independent Speaker Verification System Using Ant Colony Optimization Algorithm	421
<i>Shahla Nemati, Reza Boostani, and Mohammad Davarpanah Jazi</i>	
Image Retrieval: Color and Texture Combining Based on Query-Image	430
<i>Ilya Markov and Natalia Vassilieva</i>	
Markerless Outdoor Localisation Based on SIFT Descriptors for Mobile Applications	439
<i>Frank Lorenz Wendt, Stéphane Bres, Bruno Tellez, and Robert Laurini</i>	

Document Processing

Off-Line Arabic Handwritten Characters Recognition Based on a Hidden Markov Models	447
<i>M. Amrouche, M. Elyassa, A. Rachidi, and D. Mammass</i>	
Watermarking Based on the Density Coefficients of Faber-Schauder Wavelets	455
<i>Mohamed El hajji, Hassin Douzi, and Rachid Harba</i>	
Arabic Handwritten Characters Classification Using Learning Vector Quantization Algorithm	463
<i>Mohamed A. Ali</i>	

A Complete Pyramidal Geometrical Scheme for Text Based Image Description and Retrieval	471
<i>Guillaume Joutel, Véronique Eglin, and Hubert Emptoz</i>	
Local Orientation Extraction for Wordspotting in Syriac Manuscripts...	481
<i>P. Bilane, S. Bres, and H. Emptoz</i>	

Signal Processing

A Procedure for Efficient Generation of $1/f^\beta$ Noise Sequences	490
<i>Youcef Ferdi and Abdelmalik Taleb-Ahmed</i>	
Design and Implementation of an Image CoProcessor	498
<i>R. Ebrahimi Atani, S. Mirzakuchaki, and S. Ebrahimi Atani</i>	
Wavelet-Packet Identification of Dynamic Systems with Coloured Measurement Noise	508
<i>Henrique Mohallem Paiva and Roberto Kawakami Harrop Galvão</i>	
Comparison of the Cross Deleted Wigner Representation and the Matching Pursuit Distribution (Via Adaptive Signal Decomposition) ...	516
<i>S. Ghofrani and D.C. McLernnon</i>	

XVI Table of Contents

Mean Weight Behavior of Coupled LMS Adaptive Systems Applied to Acoustic Feedback Cancellation in Hearing Aids	527
<i>Yasmín Montenegro M. and José C.M. Bermudez</i>	
Sound Localization Based on Excitation Source Information for Intelligent Home Service Robots	536
<i>Keun-Chang Kwak</i>	
Voicing Detection in Noisy Speech Signal	544
<i>Aïcha Bouzid and Noureddine Ellouze</i>	
Coherent Synchronization of a Chaotic Communication System Based on a QAM-4 Transmitter	552
<i>Roxana Burghela, Jean Philippe Mangeot, Frédéric Launay, Patrick Coirault, and Safwan El Assad</i>	
Joint Multiple Target Tracking and Classification Using Controlled Based Cheap JPDA-Multiple Model Particle Filter in Cluttered Environment	562
<i>Zahir Messaoudi, Abdelaziz Ouldali, and Mourad Oussalah</i>	
Correlation, Independance and Inverse Modeling	570
<i>V. Vigneron</i>	
A New Approach for Reducing Embedding Noise in Multiple Bit Plane Steganography	580
<i>Arijit Sur, Piyush Goel, and Jayanta Mukhopadhyay</i>	
Region-Based Wavelet-Packet Adaptive Algorithm for Sparse Response Identification	588
<i>Odair A. Noskoski and José C.M. Bermudez</i>	
Simulation Method of SAR Raw Echo for Urban Scene	596
<i>Yingxi Zheng, Bo Zhou, Zhulin Zhong, and Ming Lv</i>	
Underwater Passive Location Technology Using Dummy Time-Reversal Mirror	605
<i>Xueli Sheng, Xizhong Bao, Junying Hui, and Guolong Liang</i>	
Multifractal Analysis on the Sphere	613
<i>Emilie Koenig and Pierre Chainais</i>	
Erratum	
Retraction Note to: Chapters	C1
<i>Abderrahim Elmoataz, Olivier Lezoray, Fathallah Nouboud, and Driss Mammass</i>	
Author Index	623

Discrete Pulse Transform of Images

Roumen Anguelov and Inger Fabris-Rotelli*

Department of Mathematics and Applied Mathematics
University of Pretoria
Pretoria 0002, South Africa
roumen.anguelov@up.ac.za,
inger.fabris-rotelli@up.ac.za

Abstract. The Discrete Pulse Transform (DPT) of images is defined by using a new class of LULU operators on multidimensional arrays. This transform generalizes the DPT of sequences and replicates its essential properties, e.g. total variation preservation. Furthermore, the discrete pulses in the transform capture the contrast in the original image on the boundary of their supports. Since images are perceived via the contrast between neighbour pixels, the DPT may be a convenient new tool for image analysis.

Keywords: LULU, discrete pulse transform, total variation preservation.

1 Introduction

One of the powerful ideas resulting from the theory of the LULU operators for sequences, developed by Carl Rohwer during the last two decades or so, is the Nonlinear Multiresolution Analysis based on a Discrete Pulse Transform (DPT), [8]. A DPT maps a bi-infinite sequence $\xi = (\dots, \xi_{-1}, \xi_0, \xi_1, \xi_2, \dots)$ onto an infinite vector

$$DPT(\xi) = (D_1(\xi), D_2(\xi), \dots) \quad (1)$$

where $D_n(\xi)$ is a sequence consisting of well separated discrete block pulses with support n . In a recent work by the authors the definition of the LULU operators was extended from sequences to multidimensional arrays, namely, functions on \mathbb{Z}^d , $d > 1$. This development opened the opportunity of developing the DPT for functions on \mathbb{Z}^d by using iterative application of the generalized LULU operators, similar to the one-dimensional case. In this paper we consider the DPT of images. Let us denote by $\mathcal{A}(\mathbb{Z}^2)$ the set of all functions defined on \mathbb{Z}^2 which have finite support. A grey scale image is a function $f \in \mathcal{A}(\mathbb{Z}^2)$ such that the support of f is a finite rectangular subset Ω of \mathbb{Z}^2 . The DPT of a function $f \in \mathcal{A}(\mathbb{Z}^2)$ is a vector of the form

$$DPT(f) = (D_1(f), D_2(f), \dots, D_N(f)) \quad (2)$$

which is finite due to the finite support of f . Here N is the number of pixels in the image and $D_n(f) = \sum_{s=1}^{\gamma(n)} \phi_{ns}$, the functions ϕ_{ns} , $s = 1, \dots, \gamma(n)$, being discrete

pulses with support of size n , $n = 1, \dots, N$. In this context a discrete pulse is a function $\phi \in \mathcal{A}(\mathbb{Z}^2)$ which is zero everywhere except on a connected set V , where it is constant. The set V is called the support of the pulse ϕ and we write $V = \text{supp}(\phi)$. The value of ϕ on V is shortly called the value of the pulse. If the value of ϕ is positive then ϕ is an up-pulse, if it is negative, ϕ is a down-pulse. Via the DPT we represent a function $f \in \mathcal{A}(\mathbb{Z}^2)$ as a sum of pulses

$$f = \sum_{n=1}^N D_n(f) = \sum_{n=1}^N \sum_{s=1}^{\gamma(n)} \phi_{ns}, \quad (3)$$

where the supports of the pulses are either disjoint or nested. Furthermore, we show that, similar to the DPT of sequences, the decomposition (3) preserves the total variation of f , that is, we have

$$TV(f) = \sum_{n=1}^N TV(D_n(f)) = \sum_{n=1}^N \sum_{s=1}^{\gamma(n)} TV(\phi_{ns}). \quad (4)$$

The total variation is an important characteristic of an image and has been successfully used in noise removal procedures, e.g. [3], [13]. The equalities in (4) show that the decomposition (3) does not create artificial variation, i.e. noise. We should further remark that the pulses in (3) have also a more direct meaning, namely, the contrast in the original image at the boundary of the support of any pulse is at least as much as the value of that pulse.

2 LULU Theory

The LULU operators on multidimensional arrays are defined in [1] by using the morphological concept of connection, [15][16], which characterizes the connected subsets of the domain. The set of connected sets appropriate for images is usually defined via a neighbor relation on the domain. More precisely, a set $C \subseteq \mathbb{Z}^2$ is called *connected* if for any two pixels $p, q \in C$ there exists a set of pixels $\{p_1, p_2, \dots, p_k\} \subseteq C$ such that each pixel is neighbour to the next one, p is neighbour to p_1 and p_k is neighbour to q . It was shown in [1] that if the neighbor relation is reflexive, symmetric, shift invariant and includes the pairs of consecutive points both vertically and horizontally, then the collection \mathcal{C} of all connected sets on \mathbb{Z}^2 is a morphological connection suitable for the LULU operators. The neighbour relations typically considered for images, e.g. as in Figure 1, satisfy these basic requirements.

Given a point $x \in \mathbb{Z}^2$ and $n \in \mathbb{N}$ we denote by $\mathcal{N}_n(x)$ the set of all connected sets of size $n + 1$ which contain a point x , that is,

$$\mathcal{N}_n(x) = \{V \in \mathcal{C} : x \in V, \text{ card}(V) = n + 1\}. \quad (5)$$

Then the basic LULU operators are defined as follows.

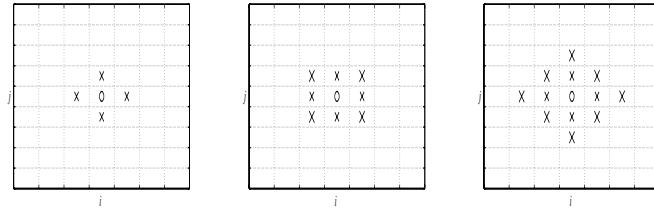


Fig. 1. Neighbours of the pixel (i, j)

Definition 1. Let $f \in \mathcal{A}(\mathbb{Z}^2)$ and $n \in \mathbb{N}$. Then

$$L_n(f)(x) = \max_{V \in \mathcal{N}_n(x)} \min_{y \in V} f(y), \quad x \in \mathbb{Z}^2, \quad (6)$$

$$U_n(f)(x) = \min_{V \in \mathcal{N}_n(x)} \max_{y \in V} f(y), \quad x \in \mathbb{Z}^2. \quad (7)$$

The operators in Definition 1 were given in [1] as a generalization of the operators L_n and U_n for sequences. However, it should be noted that these are actually the well known morphological operators of area opening and area closing (section 4.4.2, [17]) considered in a new setting, namely, the LULU theory [8][12]. It was shown in [1] that these operators are smoothers and separators (Chapter 1, [8]). Furthermore, they form, under composition, a four element totally ordered semi-group - the so called LULU semi-group [12][11] - and we have

$$L_n \leq U_n \circ L_n \leq L_n \circ U_n \leq U_n. \quad (8)$$

The smoothing effect of the operators, L_n , U_n , on a function $f \in \mathcal{A}(\mathbb{Z}^2)$ is described in terms of the concepts of local minimum and local maximum sets.

Definition 2. Let $V \in \mathcal{C}$. A point $x \notin V$ is called **adjacent** to V if $V \cup \{x\} \in \mathcal{C}$. The set of all points adjacent to V is denoted by $\text{adj}(V)$.

Definition 3. A connected subset V of \mathbb{Z}^2 is called a **local maximum set** of $f \in \mathcal{A}(\mathbb{Z}^d)$ if

$$\max_{y \in \text{adj}(V)} f(y) < \min_{x \in V} f(x).$$

Similarly V is a **local minimum set** if

$$\min_{y \in \text{adj}(V)} f(y) > \max_{x \in V} f(x).$$

Then the action of L_n and U_n on $f \in \mathcal{A}(\mathbb{Z}^2)$ can be described in terms of the following properties, [1]:

- A1 The application of L_n (U_n) removes local maximum (minimum) sets of size smaller or equal to n .

- A2 The operator L_n (U_n) does not affect the local minimum (maximum) sets directly in the sense that such sets may be affected only as a result of the removal of local maximum (minimum) sets. However, no new local minimum sets are created where there were none. This does not exclude the possibility that the action of L_n (U_n) may enlarge existing local maximum (minimum) sets or join two or more local maximum (minimum) sets of f into one local maximum (minimum) set of $L_n(f)$ ($U_n(f)$).
- A3 $L_n(f) = f$ ($U_n(f) = f$) if and only if f does not have local maximum (minimum) sets of size n or less.

In analogy with the respective concept for sequences we introduce the concept of n -monotonicity for functions in $\mathcal{A}(\mathbb{Z}^2)$.

Definition 4. Function $f \in \mathcal{A}(\mathbb{Z}^2)$ is called (locally) n -monotone if the size of all local minimum or local maximum sets of f is larger than n .

It follows from A1–A3 that for every $f \in \mathcal{A}(\mathbb{Z}^2)$

$$(L_n \circ U_n)(f) \text{ is } n\text{-monotone} \quad (9)$$

$$(U_n \circ L_n)(f) \text{ is } n\text{-monotone} \quad (10)$$

$$(L_n \circ U_n)(f) = (U_n \circ L_n)(f) = f \iff f \text{ is } n\text{-monotone} \quad (11)$$

3 Discrete Pulse Transform

The Discrete Pulse Transform (2) for a function $f \in \mathcal{A}(\mathbb{Z}^2)$ is obtained via iterative application of the operators L_n , U_n with n increasing from 1 to N . For a given n , the sequencing of L_n and U_n does not affect the properties of (2) stated in the Introduction. However, it introduces bias towards up-pulses or down-pulses. This issue will be given particular attention later. For now let P_n denote either the composition $L_n \circ U_n$ or the composition $U_n \circ L_n$ and let $Q_n = P_n \circ P_{n-1} \circ \dots \circ P_2 \circ P_1$. In the general theory of Mathematical Morphology, Q_n is known as an alternating sequential filter (see Section 8.3 [17]). However, here are interested in the portions of the image which are filtered out by the application of P_n , $n = 1, 2, \dots, N$. Indeed, we ultimately obtain $Q_N(f)$, which is a constant function containing no information about the original image except possibly the general level of illumination. The rest of the information carried by f is in the layers peeled off. More precisely, we have

$$\begin{aligned} f &= (id - P_1)(f) + ((id - P_2) \circ Q_1)(f) + ((id - P_3) \circ Q_2)(f) + \dots \\ &\quad + ((id - P_{N-1}) \circ Q_{N-2})(f) + ((id - P_N) \circ Q_{N-1})(f) + Q_N(f), \end{aligned} \quad (12)$$

where id denotes the identity operator. Let us note that a similar iterative application of area opening and area closing operators is used by Acton and Mukherjee for image classification. In [2], filtering is for selected values of n and instead of the layers of peeled off portions the authors keep record of filtered images at every scale (this would be $Q_n(f)$ in the notation adopted here). While the approach

in [2] is adequate for the classification methods derived there, the decomposition (12) contains much more information. It is also easy to apply, not least, due to the fact that it gives a representation of f in terms of discrete pulses. Indeed, we show next that (12) is a DPT decomposition of the form (3) by proving that each term is a sum of pulses of certain size. This result is obtained in the next two theorems.

Theorem 1. *Let $f \in \mathcal{A}(\mathbb{Z}^2)$ be $(n - 1)$ -monotone. Then*

- a) $(id - L_n)(f)$ is a sum of disjoint up-pulses of size n ;
- b) $(id - U_n)(f)$ is a sum of disjoint down-pulses of size n ;
- c) $(id - U_n \circ L_n)(f)$ is a sum of disjoint pulses of size n ;
- d) $(id - L_n \circ U_n)(f)$ is a sum of disjoint pulses of size n .

Proof. a) Consider $x \in \mathbb{Z}^2$ such that $(id - L_n)(f)(x) > 0$. By property A1 there exists a local maximum set V of f such that $x \in V$ and $\text{card}(V) \leq n$. Since f is $(n - 1)$ -monotone it does not have local maximum sets of size smaller than n . Therefore $\text{card}(V) = n$ and f is constant on V . Moreover, we have $(id - L_n)(f)(y) = 0$, $y \in \text{adj}(V)$. Indeed, if $(id - L_n)(f)(y) > 0$ for some $y \in \text{adj}(V)$ then y belongs to a local maximum set W of f and $\text{card}(W) \leq n$. However, any local maximum set containing y must contain V as well. Thus, $\text{card}(W) \geq n + 1$. This is a contradiction. So we have that the support of $(id - L_n)(f)$ is a union of disjoint connected sets of size n , i.e.

$$\text{supp}((id - L_n)(f)) = V_1 \cup V_2 \cup \dots \cup V_{\rho(n)},$$

where $V_s \in \mathcal{C}$ and $\text{card}(V_s) = n$ for $s = 1, 2, \dots, \rho(n)$ and $V_{s_1} \cap V_{s_2} = \emptyset$ for $s_1 \neq s_2$. Let $(id - L_n)(f)(x) = \alpha_s$ for $x \in V_s$. Then

$$(id - L_n)(f) = \sum_{s=1}^{\rho(n)} \phi_{ns},$$

where ϕ_{ns} is pulse with support V_s and value α_s .

The proof of b) is similar. The proof of c) follows from a) and b) by using that $id - L_n \circ U_n = id - L_n + (id - U_n) \circ L_n$. Similar argument holds for d).

Theorem 2. *Let the operators P_n and Q_n be as in (12). Then for every $n = 1, \dots, N$ we have that $D_n(f) = ((id - P_n) \circ Q_{n-1})(f)$ is a sum of pulses of the form*

$$D_n(f) = \sum_{s=1}^{\gamma(n)} \phi_{ns},$$

where $\text{card}(\text{supp}(\phi_{ns})) = n$ and $\text{supp}(\phi_{ns_1}) \cap \text{supp}(\phi_{ns_2}) = \emptyset$ for $s_1 \neq s_2$. Moreover, the supports of all pulses ϕ_{ns} , $s = 1, \dots, \gamma(n)$, $n = 1, \dots, N$ are either disjoint or nested, that is,

$$(\text{supp}(\phi_{n_1 s_1}) \cap \text{supp}(\phi_{n_2 s_2}) \neq \emptyset, n_1 \leq n_2) \implies \text{supp}(\phi_{n_1 s_1}) \subseteq \text{supp}(\phi_{n_2 s_2}). \quad (13)$$

Proof. Let $n \in \{2, \dots, N\}$. The function $Q_{n-1}(f) = P_{n-1}(Q_{n-2}(f))$ is $(n-1)$ -monotone due to (9)–(10). Then it follows from Theorem II that $(id - P_n)Q_{n-1}(f)$ is a sum of pulses of size n . For $n = 1$ the function f is trivially 1-monotone and hence $(id - P_1)(f)$ is a sum of pulses of size 1. Note that here we may have $P_n = U_n \circ L_n$ and obtain the result by item c) in Theorem II, or use item d) if $P_n = L_n \circ U_n$. Due to (9)–(10) the sequencing of L_k and U_k in the operator P_k for $k < n$ also does not matter. Hence it remains to prove (13).

Let $\text{supp}(\phi_{n_1 s_1}) \cap \text{supp}(\phi_{n_2 s_2}) \neq \emptyset$. It follows from the construction of the pulses in the proof of Theorem II that the functions $Q_n(f)$ and $P_{n+1}(Q_n(f))$, $n \geq n_1$, are constants on the set $\text{supp}(\phi_{n_1 s_1})$. Furthermore, the set $\text{supp}(\phi_{n_2 s_2})$ is a local maximum set or a local minimum set of either $Q_{n_2-1}(f)$ or $P_{n_2}(Q_{n_2-1}(f))$. Then by the definition of local maximum and local minimum sets it follows that $\text{supp}(\phi_{n_1 s_1}) \subset \text{supp}(\phi_{n_2 s_2})$.

As mentioned earlier the sequencing of the operators L_n and U_n in P_n results in bias towards up-pulses or down-pulses. The operator F_n , called a *floor operator* given by $F_n = P_n \circ P_{n-1} \circ \dots \circ P_2 \circ P_1$ where $P_n = U_n \circ L_n$, is biased towards up-pulse. This means that for any particular size the up-pulses are extracted before the down pulses. Similarly the operator C_n , called a *ceiling operator* given by $C_n = P_n \circ P_{n-1} \circ \dots \circ P_2 \circ P_1$ where $P_n = L_n \circ U_n$, is biased towards down-pulses. A DPT using F_n will generally have more up-pulses and fewer down-pulses compared to DPT using C_n . It is important to realize that the bias extremes will differ for different original images due to their size and contrast so interpretation should be relative to the current image.

The alternating bias separators Z_n^+ and Z_n^- (see [4] and references therein) are given by:

$$Z_{n+1}^- = \begin{cases} L_{n+1}U_{n+1}Z_n^- & \text{if } n \text{ is even} \\ U_{n+1}L_{n+1}Z_n^- & \text{if } n \text{ is odd} \end{cases}$$

$$Z_{n+1}^+ = \begin{cases} L_{n+1}U_{n+1}Z_n^+ & \text{if } n \text{ is odd} \\ U_{n+1}L_{n+1}Z_n^+ & \text{if } n \text{ is even} \end{cases}$$

Using the order in the LULU semi-group (8) it is easy to see that

$$F_n \leq Z_n^\pm \leq C_n .$$

We should note that if $f \in \mathcal{A}(\mathbb{Z}^2)$ is such that $F_n = C_n$, $n = 1, \dots, N$, then f has a unique DPT decomposition of the form (12).

Finally we discuss the total variation preservation property (4). There are several different concepts used for total variation of a function of multidimensional argument in literature. For a function $f \in \mathcal{A}(\mathbb{Z}^2)$ we define the total variation as

$$TV(f) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} (|f(i+1, j) - f(i, j)| + |f(i, j+1) - f(i, j)|) . \quad (14)$$

It was proved in [1] that the operators L_n and U_n are total variation preserving. Furthermore, the composition of total variation preserving operators is also total variation preserving. Therefore the decomposition (3) obtained via (12) preserves the total variation, i.e. (4) holds.

Table 1. Bias Percentages ($\times 10^{-3}$) between F_n, C_n, Z_n^+ and Z_n^-

	C_n	F_n	Z_n^+	Z_n^-
C_n	0	0.2609	0.2269	0.0340
F_n	0.2609	0	0.0340	0.2269
Z_n^+	0.2269	0.0340	0	0.2610
Z_n^-	0.0340	0.2269	0.2610	0

The decomposition of the original image can be done using F_n, C_n, Z_n^+ or Z_n^- . While it is possible that these lead to different results it is seldom significantly observable in practice. For example, the percentage bias between the operators, when applied to the image (a) in Figure 2, is given in Table 1.

4 Illustrative Examples

4.1 A Partial Reconstruction

Consider image (a) in Figure 2. We wish to extract the boat as a feature and remove the background water. We consider any feature not part of the boat as



(a)



(b)



(c)



(d)

Fig. 2. (a) Original image; (b) Pulses of size up to 80; (c) Pulses of size from 81 to 4000; (d) Pulses of size larger than 4000

noise. The images (b) - (d) in Figure 2 are partial reconstructions, see [1], of the original. We can consider (b) as the microfeatures e.g. noise, (c) as the medium-sized features and (d) as the large features. We see that pulses of size up to 80 represent the water, pulses from 81 to 4000 the boat and pulses larger than 4000 represent larger noise patterns.

4.2 Identifying Pulses of Interest

In Figure 3 the pulses of size 100 to 300 with relative luminosity greater than 1 represent the solar glint on the waves (see image (b) in Figure 3) and they are emphasized in image (c) by giving them all black borders. So we are able to extract the small solar glint patches from the image and can subsequently reconstruct the image without them. The pulses of size 1501 to 2500 are the pulses corresponding to the boat (as seen in image (d) of Figure 3).



(a)



(b)



(c)



(d)

Fig. 3. (a) Original image; (b) Pulses of size 100 to 300 with relative luminosity greater than 1; (c) Borders of the pulses of size 100 to 300 with relative luminosity greater than 1 superimposed on the original image; (d) Pulses of size larger than 1501 to 2500

References

1. Anguelov, R., Plaskitt, I.: A Class of LULU Operators on Multi-Dimensional Arrays, Technical Report UPWT2007/22, University of Pretoria (2008), <http://arxiv.org/abs/0712.2923>
2. Acton, S.T., Mukherjee, D.P.: Scale Space Classification Using Area Morphology. IEEE Transactions on Image Processing 9(4) (2000)
3. Chambolle, A., Lions, P.-L.: Image recovery via total variation minimization and related problems. Numerische Mathematik 76, 167–188 (1997)
4. du Toit, J.P.: The Discrete Pulse Transform and Applications, Masters Thesis, Univeristy of Stellenbosch (2007)
5. Epp, S.S.: Discrete Mathematics with Applications. International Thomson Publishing (1995)
6. Laurie, D.P., Rohwer, C.H.: The discrete pulse transform. SIAM J. Math. Anal. 38(3) (2007)
7. Matheron, G.: Filters and lattices, in Image Analysis and Mathematical Morphology. In: Serra, J. (ed.) Theoretical Advances, ch. 6, vol. II, pp. 115–140. Academic Press, London (1988)
8. Rohwer, C.H.: Nonlinear Smoothing and Multiresolution Analysis, Birkhäuser (2000)
9. Rohwer, C.H.: Variation reduction and LULU-smoothing. Quaestiones Mathematicae 25, 163–176 (2002)
10. Rohwer, C.H.: Fully trend preserving operators. Quaestiones Mathematicae 27, 217–230 (2004)
11. Rohwer, C.H., Wild, M.: Natural Alternatives for One Dimensional Median Filtering. Quaestiones Mathematicae 25, 135–162 (2002)
12. Rohwer, C.H., Wild, M.: *LULU* Theory, Idempotent Stack Filters, and the Mathematics of Vision of Marr. Advances in imaging and electron physics 146, 57–162 (2007)
13. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D 60, 259–268 (1992)
14. Serra, J.: Image Analysis and Mathematical Morphology. Academic Press, London (1982)
15. Serra, J.: Image Analysis and Mathematical Morphology. In: Serra, J. (ed.) Theoretical Advances, vol. II. Academic Press, London (1988)
16. Serra, J.: A lattice approach to image segmentation. Journal of Mathematical Imaging and Vision 24, 83–130 (2006)
17. Soille, P.: Morphological Image Analysis. Springer, Heidelberg (1999)

Constrained Unsharp Masking for Image Enhancement

Radu Ciprian Bilcu and Markku Vehvilainen

Nokia Research Center, Visiokatu 1, 33720, Tampere, Finland

radu.bilcu@nokia.com, markku.vehvilainen@nokia.com

<http://research.nokia.com/~bilcu>

Abstract. In this paper we present a cost-effective solution for combined de-noising and sharpening of digital images. Our method combines the unsharp masking and sigma filtering techniques through a regularization mechanism thus ensuring effective noise reduction and edge enhancement in the processed image. We describe our method in detail and we analyze the proposed implementation through extensive experiments done in various scenarios. Due to its low computational complexity the proposed method is well suited for mobile implementations.

Keywords: Image sharpening, de-noising, unsharp masking.

1 Introduction

Digital photography has gained an increased interest, in the mobile devices industry, during the last decade. It is a common fact nowadays that mobile devices, such as mobile phones and PDA's, are equipped with digital cameras able to capture images and video clips. Due to their specific hardware limitations the quality of the captured images are not yet at the level of the SLR (single-lens reflex) cameras. The resolution of the imaging sensors, used in the cameraphones increases, but the total sensor size is limited, which decreases the active area of a sensor pixel. As a consequence, the captured images have a lower signal to noise ratio. Miniaturization and cost limitations enforce the use of lower quality optical systems in the cameraphones compared with the ones used in the SLR cameras. This very often introduces blur in the captured images and video clips. The low signal to noise ratio and the optical blur are probably the most important distortions that must be corrected by an imaging chain designed to be implemented in cameraphones. A very large amount of research has been done in the field of image de-noising and sharpening in order to address these two important problems and various solutions have been proposed in the open literature. Mobile devices possess a very limited processing power and contain several applications that must run in parallel which decreases even more the resources allocated for image processing tasks. As a consequence, there is a large interest in the development of low complexity image processing solutions suitable for mobile implementations.

In this paper we propose a cost effective solution to the problem of combined de-noising and sharpening of digital images. Our proposed method combines the sigma filtering and unsharp masking (USM), through a simple regularization step, to accomplish both noise reduction and edge enhancement. The reason why we have selected these two techniques is due to their relatively good performances obtained at a very low computational cost. The paper is organized as follows: in Section 2 several related existing sharpening methods are briefly reviewed. In Section 3 we introduce our proposed approach and we illustrate its functionality through simple experiments done on synthetic images. In Section 4 we show the results of our method obtained for real images and Section 5 concludes the paper.

2 Existing Approaches

The problem of image sharpening has been extensively studied and many approaches have been proposed in the open literature. Bilateral filter have been used in [13] to implement a combined method for sharpening and de-noising. The novelty of this method consists in the introduction of an offset term in the range filter part of the bilateral filter. This offset is adapted to the local characteristics of the input image such that the behavior of the algorithm changes from smoothing to edge enhancement according to the local characteristics of the image. Although its good performances, this algorithm needs some optimization of the parameters which is done by training on a set of images. Image sharpening has been implemented also by means of fuzzy networks [10], using methods modeling the human visual system [6], by non-linear filtering in transform domain [7] or using diffusion filtering [3], to mention a few. Although their good performance, the above mentioned approaches have some characteristics that makes them to be quite difficult to be implemented in low power devices. For instance the method in [10] make use of 2 fuzzy networks connected in parallel and have longer processing time compared with other approaches. The nonlinear methods from [6] and [7], despite their excellent performance, have also relatively large computational complexity. The coupled shock-diffusion filtering method from [3] is expected to have also a long processing time since it is an iterative process.

Despite its simplicity, the unsharp masking technique [9] provides relatively good performance in enhancing the image sharpness thus representing a good alternative for practical implementations. In the unsharp masking technique the high frequency content of the input image is emphasized as depicted in Fig. 7. According to this figure the input image is filtered by a high-pass filter (the dotted block denoted as HPF). The high-pass filtered image $I_H(i, j)$ is then multiplied by a constant parameter λ and added to the input image. There are several possibilities to implement the HPF filter. We should mention here that Fig. 7 suggests to implement the HPF filtering as the difference between the input image and its smoothed version (the output of the low-pass filter LPF). Equally well this operation can be implemented by a single high-pass convolution kernel [9]. Despite its simplicity and good performance, in many applications, the unsharp masking technique has two main drawbacks. First of all the noise present

in the input image is amplified due to the presence of the high-pass filter. While this might not be visually disturbing in image regions with high activity (large number of edges), on smooth image areas noise amplification can give unpleasant visual distortions. Another disadvantage of the unsharp masking technique is the presence of the overshooting artifacts. This is due to a too large amplification of the high contrast areas. Among several solutions, proposed to eliminate or at least reduce these two disadvantages, the adaptive unsharp masking shown to be very effective [9]. The main idea of this method is to use a variable parameter λ which is adaptive to the image local contrast. In this manner the regions with medium contrast are enhanced more while the smooth regions and those with high contrast are less enhanced. However, in [9], modification of λ is done by means of an adaptive filter which implies matrix inversion and increased number of parameters to setup.

Another way to reduce the overshooting in the unsharp masking is to introduce clipping of the overshoots to the nearest local extremes [11], [8]. Some of the techniques that integrates clipping necessitate edge detection in order to be able to locate the position of the overshoots. In this paper we show that clipping not only the edge pixels, but every pixels from the image, one could obtain both de-noising and edge sharpening. We propose a modified processing scheme that combines sigma smoothing, unsharp masking and clipping to reach our goals.

3 Proposed Approach

In this section we describe our proposed method for combined de-noising and sharpening. In order to introduce our method we start from the block diagram

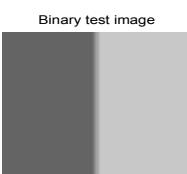


Fig. 1. Test image

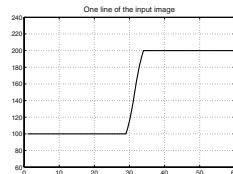


Fig. 2. One line from test image

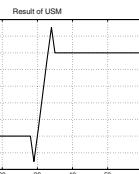


Fig. 3. Enhanced by USM

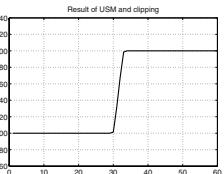


Fig. 4. Result after clipping

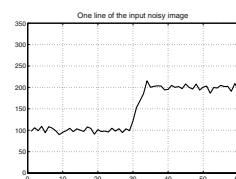


Fig. 5. One noisy line from test image

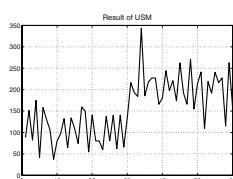


Fig. 6. Enhanced by USM

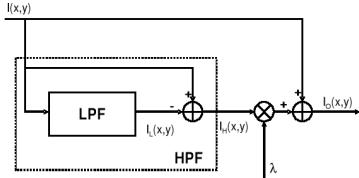


Fig. 7. Block diagram of the unsharp masking

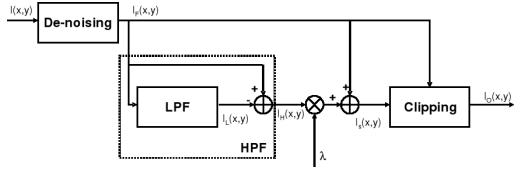


Fig. 8. Block diagram of the cascaded de-noising and sharpening with clipping

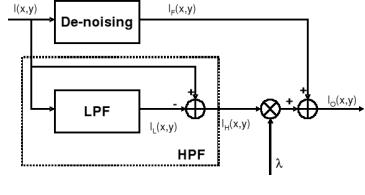


Fig. 9. Block diagram of the integrated de-noising and sharpening without clipping

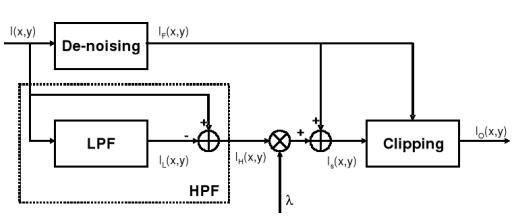


Fig. 10. Block diagram of integrated de-noising and sharpening with clipping

of the USM depicted in Fig. 7. Due to the presence of the high-pass filter (HPF) the processed image $I_O(i, j)$ might contain overshooting distortions around sharp edges. This effect is illustrated in Fig. 2 and Fig. 3. In Fig. 2 one line from the binary image of Fig. 1 is shown and in Fig. 3 the result of USM is depicted. These overshooting distortions can be eliminated by clipping them to the extreme values from their neighborhood as shown in Fig. 4. Now lets look to the case of noisy input images. To this end in Fig. 5 the same line of the binary image, with added zero mean Gaussian noise, is shown. In Fig. 6 we show the enhanced output using USM. One can notice, from this figure, the overshooting and noise amplification effects. Analyzing the plots from Fig. 5 and Fig. 6 we can observe that clipping would not be very effective in the noisy case. Indeed, the local extremes around overshoots are affected by the input noise and clipping them using the noisy local extremes could introduce artifacts. From this observation we can conclude that clipping would be more effective if it is made according to a noise-free version of the input image instead of the original noisy one. As a consequence we modify the block diagram of the USM with clipping as shown in Fig. 10. Moreover instead of clipping only edge pixels we perform clipping for all image pixels. Clipping is done according to the smoothed image $I_F(i, j)$. In this manner in smooth areas the noise level from the processed image is limited to the level of the remaining noise in $I_F(i, j)$. As a consequence, the noise in the output image will be at most equal with the noise level from $I_F(i, j)$ and the edges of $I_O(i, j)$ will be enhanced without overshooting.



Fig. 11. From left to right and up to bottom (part of the images): Original image, noisy blurred input image, the result of the implementation from Fig. 10, the result of the implementation from Fig. 9, the result of USM (Fig. 7), the result of the implementation from Fig. 8

According to Fig. 10, the flow of our proposed algorithm is as follows:

1. Compute the image $I_F(i, j)$ by smoothing the input image $I(i, j)$. The smoothing process should retain as many image details as possible while reducing the input noise. Due to its simple implementation we have selected for this step the sigma filtering method [4]. Other more complex image de-noising methods can be implemented as well if there is enough computational power available (see for instance [1] and the references therein).
2. Compute the image $I_H(i, j)$ by high-pass filtering of the input image $I(i, j)$. This can be done by convolving $I(i, j)$ with a high pass kernel $g(k, l)$.

$$I_H(i, j) = (I \otimes g)(i, j) \quad (1)$$

Another alternative to obtain $I_H(i, j)$ is to compute it as the difference between the original image $I(i, j)$ and its low-pass version $I_L(i, j)$ as follows:

$$I_H(i, j) = I(i, j) - I_L(i, j), \quad I_L(i, j) = (I \otimes h)(i, j). \quad (2)$$

with $h(i, j)$ being a low pass filtering kernel.

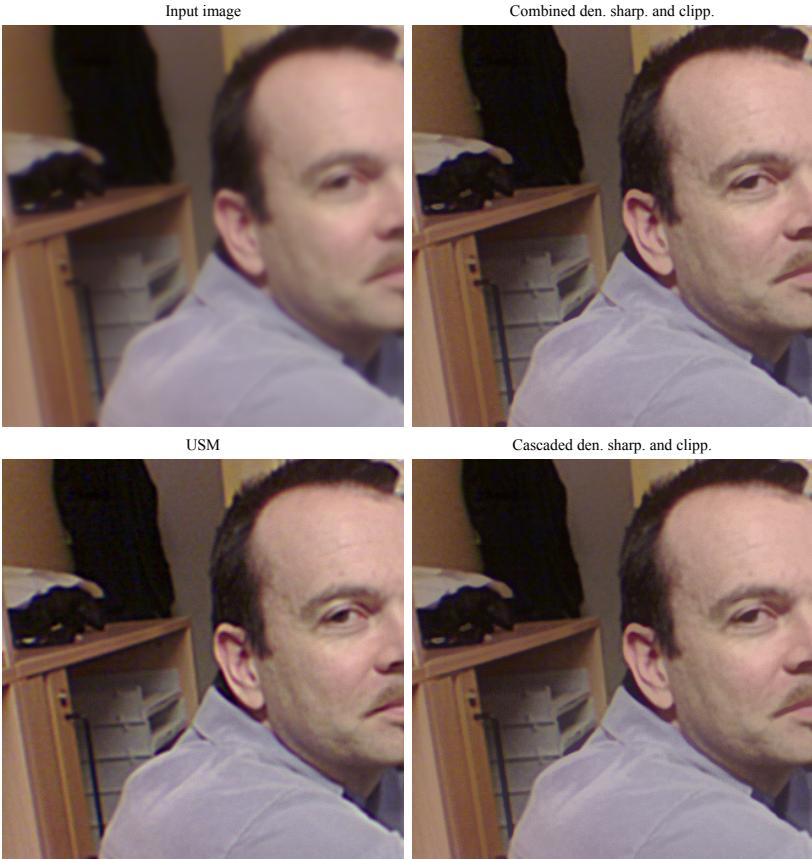


Fig. 12. From left to right and up to bottom (part of the images): original image captured by a cameraphone, the result of the implementation from Fig. 10, the result of USM (Fig. 7), the result of the method shown in Fig. 8

3. Add the enhanced high-pass image $I_H(i, j)$ to the de-noised image $I_F(i, j)$:

$$I_S(i, j) = I_F(i, j) + \lambda I_H(i, j). \quad (3)$$

- where λ is a constant parameter controlling the strength of sharpening.
4. Clipping the enhanced image $I_S(i, j)$ according to the local extremes from $I_F(i, j)$:

$$I_O(i, j) = \begin{cases} I_S(i, j) & \text{if } I_S(i, j) \in [MIN, MAX], \\ MIN & \text{if } I_S(i, j) < MIN, \\ MAX & \text{if } I_S(i, j) > MAX. \end{cases} \quad (4)$$



Fig. 13. Original image captured by a cameraphone (left column) and the result of the implementation from Fig. 10 (right column)

The MIN and MAX values are obtained from a $T \times T$ rectangular window $\Omega_{T \times T}$ centered at position i, j in $I_F(i, j)$.

Another alternative implementation is depicted in Fig. 8. However there are differences between the method illustrated in Fig. 10 and the method depicted in Fig. 8. In the first method the high-pass filtered image $I_H(i, j)$ is computed from the input image while in Fig. 8 it is computed from the de-noised image $I_F(i, j)$. De-noising will smooth also some details of the image such that cascading de-noising and USM as in Fig. 8 those details might not be enough enhanced. On the contrary, the details that would be smoothed in the de-noising process are still enhanced in Fig. 10 due to the fact that $I_H(i, j)$ is obtained from the original image. Since the output $I_O(i, j)$ is obtained by adding the enhanced high frequency content to the de-noised image $I_F(i, j)$, in both methods, some under enhancement of small details can be expected. However, in implementation shown in Fig. 10 this effect is somehow reduced due to the fact that $I_H(i, j)$ is computed from the input image. Anyway, both implementations have the same computational complexity and they perform similar in most of the cases.

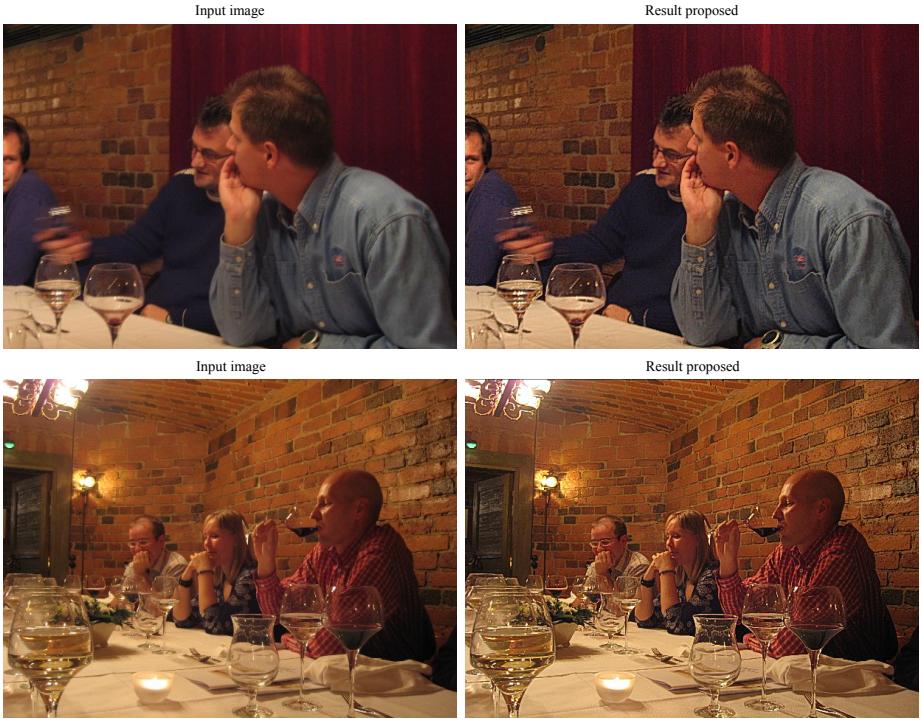


Fig. 14. Original image captured by a cameraphone (left column) and the result of the implementation from Fig. 10 (right column)

4 Experiments and Results

In this section we present experimental results showing the performance of our proposed approach on gray scale and color digital images. We have performed extensive experiments to verify the behavior of our new method on several well known test images as well as on images captured with cameraphones and dedicated digital cameras. The image data-base, which we have used in our experiments, contains more than 80 gray scale and color images captured with different imaging devices.

In the first set of experiments, presented here, the 8 bit image *Lena* was first blurred using a Gaussian kernel with $\sigma_G^2 = 1$ and then a zero mean Gaussian distributed noise with variance $\sigma_n^2 = 25$ was added to it. In Fig. 11 the processed images by the methods described in Fig. 7, Fig. 8, Fig. 9 and Fig. 10 are shown. We can see that both implementations from Fig. 8 and Fig. 10 enhance the image sharpness without increasing the noise level. This is due to the clipping mechanism that is applied to all image pixels. As expected, the USM method and the implementation from Fig. 9 suffers from noise amplification and overshooting.

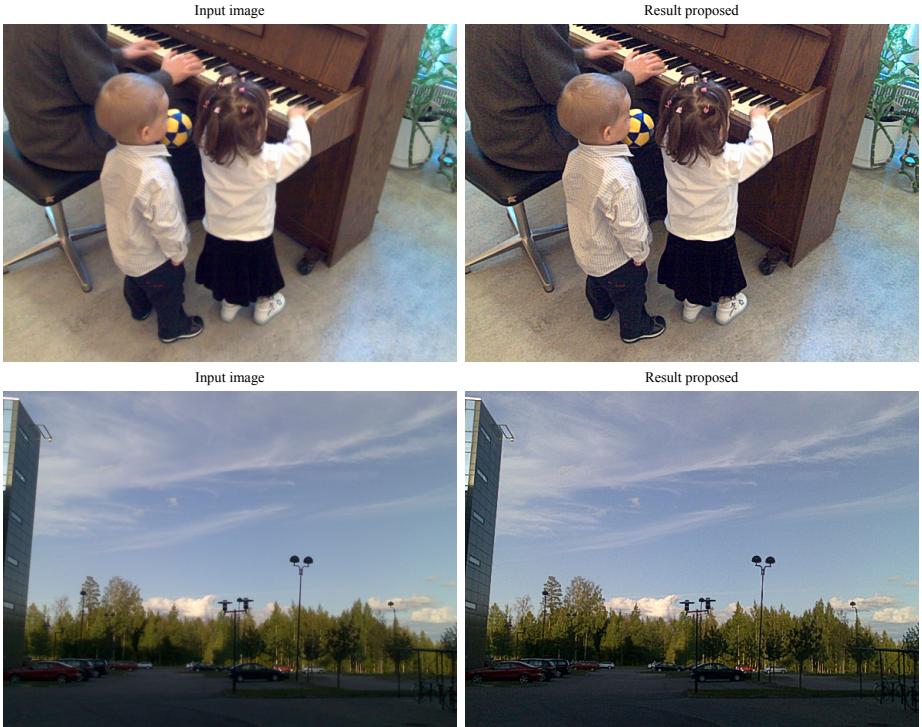


Fig. 15. Original image captured by a cameraphone (left column) and the result of the implementation from Fig. 10 (right column)

In the second set of experiments, several 8 bit digital color images, captured by different cameraphones and dedicated digital cameras, have been processed by the same enhancement methods and the results are shown in Fig. 12, Fig. 13, Fig. 14 and Fig. 15. Again we can see that our proposed methods ensure more pleasant visual results with enhanced edges, reduced noise and no overshoots also for digital color images.

The sigma filter, used in our approach, necessitates the knowledge of the noise variance. In our implementation we have used the method from [2] to estimate the noise variance.

5 Conclusions

In this paper we have introduced a noise reduction and sharpening method suitable for mobile implementations. In our method we combined the sigma de-noising and the unsharp masking techniques through a simple clipping mechanism. We have seen that if clipping is done for all image pixels and not only for over enhanced edges the noise level in the processed image can be reduced

together with increasing the edge sharpness. The proposed method can be combined with other existing approaches that adaptively adjusts the strength of sharpening based on the local image activity. Further developments in this direction as well as better clipping alternatives are currently under consideration and the results will be published elsewhere.

References

1. Bilcu, R.C., Vehvilainen, M.: A Novel Decomposition Scheme for Image De-noising. In: IEEE Int. Conf. on Ac. Speech and Sig. Proc., pp. 577–580 (2007)
2. Bilcu, R. C., Vehvilainen, M.: A New Method for Noise Estimation in Images. IEEE Int. Work. on Nonlin. Sig. and Im. Proc. (2005)
3. Fu, S., Ruan, Q., Wang, W.: A Coupled Shock-Diffusion Filter for Image Enhancement. In: IEEE Int. Conf. on Sig. Proc., pp. 16–20 (2006)
4. Lee, J.S.: Digital Image Smoothing and the Sigma Filter. Comp. Graph. Im. Proc. 24, 255–269 (1983)
5. Lukac, R., Plataniotis, K.: N.: Cost-Effective Sharpening of Single-Sensor Camera Images. In: IEEE Int. Conf. on Multimedia and Expo., pp. 829–832 (2006)
6. Matz, S.C., Figueiredo, R.J.P.: A Nonlinear Image Contrast Sharpening Approach Based on Munsell's Scale. IEEE Trans. on Im. Proc. 15(4), 900–909 (2006)
7. Mitra, K.S., Yu, T.-H.: Nonlinear Filters for Image Sharpening and Smoothing. IEEE Int. Conf. on Syst. Eng, 241–244 (1991)
8. Ojo, O., Kwaaitaal-Spassova, A., An Algorithm, T.G.: for Integrated Noise Reduction and Sharpness Enhancement. IEEE Trans. on Cons. El. 46(3), 474–480 (2000)
9. Polesel, A., Ramponi, G., Mathews, V.G.: Image Enhancement via Adaptive Unsharp Masking. IEEE Trans. on Im. Proc. 9(3), 505–510 (2000)
10. Russo, F.: An Image Enhancement Technique Combining Sharpening and Noise Reduction. IEEE Trans. on Instr. and Measur. 51(4), 824–828 (2002)
11. Rieder, P., Scheffer, G.: New Concepts on Denoising and Sharpening of Video Signals. IEEE Trans. on Cons. El. 47(3), 666–671 (2001)
12. Vehvilainen, M., Yrjanainen, J.: Circuit Arrangement to Accentuate the Sharpness and Attenuate the Noise of a Television Image. European Patent EP95101009 (1994)
13. Zhang, B., Allebach, J.: P.: Adaptive Bilateral Filter for Sharpness Enhancement and Noise Removal. In: IEEE Int. Conf. on Im. Proc., pp. VI417–VI420 (2007)
14. Wang, Q., Ward, R., Zou, J.: Contrast Enhancement for Enlarged Images Based on Edge Sharpening. IEEE Int. on Im. Proc. (2005)

Discrete Wavelet Diffusion for Image Denoising

Kashif Rajpoot¹, Nasir Rajpoot², and J. Alison Noble¹

¹ Department of Engineering Science, University of Oxford, UK

² Department of Computer Science, University of Warwick, UK

{kashif, noble}@robots.ox.ac.uk, nasir@dcs.warwick.ac.uk

Abstract. Nonlinear diffusion, proposed by Perona-Malik, is a well-known method for image denoising with edge preserving characteristics. Recently, nonlinear diffusion has been shown to be equivalent to iterative wavelet shrinkage, but only for (1) Mallat-Zhong dyadic wavelet transform and (2) Haar wavelet transform. In this paper, we generalize the equivalence of nonlinear diffusion to non-linear shrinkage in the standard discrete wavelet transform (DWT) domain. Two of the major advantages of the standard DWT are its simplicity (as compared to 1) and its potential to benefit from a greater range of orthogonal and biorthogonal filters (as compared to both 1 and 2). We also extend the wavelet diffusion implementation to multiscale. The qualitative and quantitative results shown for a variety of images contaminated with noise demonstrate the promise of the proposed standard wavelet diffusion.

1 Introduction

A common problem with image or video signals is the contamination with undesired noise. This poses problems both for visual quality and automated analysis. Many automated analysis operations (e.g., segmentation, edge detection) are highly sensitive to the noise. It is often desirable to remove or suppress the noise by means of a pre-processing operation. Many denoising methods consider noise as either high-frequency content or outlier, potentially confusing the edges with noise and smoothing them during the process of noise removal. Edges are important features containing structural information that should be preserved. Nonlinear diffusion, initially proposed by Perona-Malik [1], is a popular iterative method for noise removal that attempts to perform edge-preserving smoothing.

During the last decade or so, the discrete wavelet transform (DWT) has also been shown [2] to be a highly successful way of signal denoising, as it effectively separates signal and noise in the wavelet domain. The DWT based denoising, termed also as wavelet shrinkage, works by thresholding (often, nonlinearly) the wavelet coefficients before reconstructing the denoised signal from wavelet decomposition coefficients. One of the problems with wavelet shrinkage is that it results in valleys around edges (termed as *ringing artifacts*) which can be reduced by cycle spinning [3], at the cost of increased computational and storage complexity.

Recently, it was shown [4, 5] that a single step of nonlinear diffusion can be considered equivalent to a single shrinkage iteration of coefficients of Mallat-Zhong dyadic wavelet transform (MZ-DWT) [6]. In other words, it is possible to perform the

diffusion in the wavelet domain such that the benefits of both the techniques could be combined. Nonlinear diffusion begins with an edge estimate, typically obtained through a gradient operator, which may be badly influenced by the noise present in the image. It is believed that the edge estimate in the wavelet domain is much less sensitive to noise due to its characteristics of better low- and high-frequency separation. However, MZ-DWT is an uncommon variant of standard DWT and is not commonly utilized. It has its own subband filtering framework and a set of wavelet filters, which are usually derived from the derivative of a smoothing function (e.g., cubic spline). Moreover, MZ-DWT may not be able to benefit from the great variety of orthogonal and biorthogonal wavelet filters. In another work, the diffusivity function of nonlinear diffusion is shown [7] to have equivalence with Haar wavelet shrinkage. Again, the work is restricted only to the use of Haar wavelet filter.

In this paper, we combine the power of nonlinear diffusion and wavelet shrinkage and generalize their equivalence by establishing a relationship between the standard DWT and nonlinear diffusion. The generalization enables one to perform diffusion filtering utilizing the wavelet coefficients obtained from standard DWT, which is simple and commonly used in image processing research. Thus, one can take advantages of both popular denoising methods – wavelets and nonlinear diffusion – in a single wavelet diffusion framework. We hope that this work will make the wavelet diffusion more widely usable while offering the use of a range of wavelet filters.

This paper is organized as follows: next section establishes the relationship between DWT and nonlinear diffusion after briefly introducing them. Section 3 provides experimental details and presents some qualitative and quantitative results. The paper finishes with a brief discussion of results, conclusions and future directions.

2 Discrete Wavelet Diffusion

We present a brief overview of Perona-Malik nonlinear diffusion and DWT and later show the equivalence between the two frequently used denoising techniques.

2.1 Nonlinear Diffusion

The nonlinear diffusion filtering operation is governed by the following nonlinear partial differential equation (PDE):

$$\frac{\partial}{\partial t} [f_t(x, y)] = \operatorname{div}[c(x, y) \cdot \nabla f_t(x, y)] \quad (1)$$

where $f_t(x, y)$ is the initial noisy image at time t and the solution of the above PDE iteratively yields a filtered version which can be considered as the denoised image. In the above, $c(x, y)$ is the diffusion coefficient controlling the diffusion amount and is typically described by a diffusivity function g of the form:

$$c(x, y) = g(|\eta(x, y)|) = \frac{1}{1 + \left(\frac{|\eta(x, y)|}{\lambda}\right)^2} \quad (2)$$

where $|\eta(x, y)|$ is the edge estimate at pixel (x, y) , usually approximated by a gradient-magnitude operator, and λ is the edge threshold parameter. The diffusivity function $g(|\eta|)$ is a nonnegative monotonically decreasing function and, depending upon the value of the edge threshold λ , encourages homogenous regions with reduced noise while preserving the edges. The nonlinear diffusion PDE of (1) can be expanded to:

$$\frac{\partial}{\partial t}[f_t(x, y)] = \frac{\partial}{\partial x}\left[c(x, y) \cdot \frac{\partial}{\partial x} f_t(x, y)\right] + \frac{\partial}{\partial y}\left[c(x, y) \cdot \frac{\partial}{\partial y} f_t(x, y)\right] \quad (3)$$

substituting the time-derivative of $f_t(x, y)$ by its forward difference and discretizing:

$$\frac{f_{t+\Delta t}(x, y) - f_t(x, y)}{\Delta t} = \frac{d}{dx}\left[c(x, y) \cdot \frac{d}{dx} f_t(x, y)\right] + \frac{d}{dy}\left[c(x, y) \cdot \frac{d}{dy} f_t(x, y)\right]$$

Let $\Delta t = 1$ and replacing $c(x, y)$, as defined in (2) above, with $1 - p(x, y)$, we obtain:

$$\begin{aligned} f_{t+1}(x, y) &= f_t(x, y) + \frac{d^2}{dx^2} f_t(x, y) + \frac{d^2}{dy^2} f_t(x, y) \\ &\quad - \frac{d}{dx}\left[p(x, y) \cdot \frac{d}{dx} f_t(x, y)\right] - \frac{d}{dy}\left[p(x, y) \cdot \frac{d}{dy} f_t(x, y)\right] \end{aligned} \quad (4)$$

Usually, $f_t(x, y)$ is initialized by $f_0(x, y) = f(x, y)$, where $f(x, y)$ is the original noisy image. The solution to the above equation yields a denoised image after a certain number of iterations.

2.2 Discrete Wavelet Transform (DWT)

Wavelets offer a simple and elegant framework for simultaneous time-frequency and multiresolution analysis. In wavelet analysis, a function is decomposed into a superposition of dilated and scaled versions of a mother wavelet basis function. The decomposition highlights interesting features in the wavelet domain which are often not visible in the spatial domain. In DWT, an image is filtered into four subbands at each resolution and the low frequency subband is further filtered through an iterative process to provide the multiresolution representation. Here, we give an overview of the fast implementation of the discrete wavelet transform (DWT) [8]. The *forward wavelet transform* (FDWT) operation (termed also as *decomposition* or *analysis*) is given by:

$$\begin{aligned} A_j(x, y) &= f(x, y) \otimes H(y) \otimes H(x) \\ W_j^H(x, y) &= f(x, y) \otimes H(y) \otimes G(x) \\ W_j^V(x, y) &= f(x, y) \otimes G(y) \otimes H(x) \\ W_j^D(x, y) &= f(x, y) \otimes G(y) \otimes G(x) \end{aligned} \quad (5)$$

where $f(x, y)$ is the given image, \otimes denotes the convolution operation, and H and G are one-dimensional low-pass and high-pass decomposition filters, respectively. The image is decomposed into four subbands; A_j denotes the low-frequency approximation subband and $W_j^i, i \in \{H, V, D\}$ denotes the high-frequency subband at scale j and i being one of the horizontal (H), vertical (V), or diagonal (D) orientations.

The original image can be reconstructed through the *inverse wavelet transform* (IDWT) operation (termed also as *reconstruction* or *synthesis*):

$$\begin{aligned} f(x, y) = & A_j(x, y) \otimes \tilde{H}(x) \otimes \tilde{H}(y) + W_j^H(x, y) \otimes \tilde{G}(x) \otimes \tilde{H}(y) \\ & + W_j^V(x, y) \otimes \tilde{H}(x) \otimes \tilde{G}(y) + W_j^D(x, y) \otimes \tilde{G}(x) \otimes \tilde{G}(y) \end{aligned} \quad (6)$$

where \tilde{H} and \tilde{G} are the one-dimensional low-pass and high-pass reconstruction filters and derived as conjugate or dual of the decomposition filters H and G depending upon whether the filters are orthogonal or biorthogonal, respectively.

Assuming that the original image $f(x, y)$ is part of the scale-space such that the wavelet coefficients can be manipulated after decomposition, the original image and the reconstructed image can be considered as $f_t(x, y)$ and $f_{t+1}(x, y)$, respectively. Using this convention and writing the analysis and synthesis operations in joint form:

$$\begin{aligned} f_{t+1}(x, y) = & [f_t(x, y) \otimes HH(x, y)] \otimes \tilde{HH}(x, y) \\ & + [f_t(x, y) \otimes GH(x, y)] \otimes \tilde{GH}(x, y) \\ & + [f_t(x, y) \otimes HG(x, y)] \otimes \tilde{HG}(x, y) \\ & + [f_t(x, y) \otimes GG(x, y)] \otimes \tilde{GG}(x, y) \end{aligned} \quad (7)$$

where $f(x, y) \otimes GH(x, y)$ is the separable convolution of $f(x, y)$ with $G(x)$ and $H(y)$. Transforming the above equation into Fourier domain, we get:

$$\begin{aligned} \hat{f}_{t+1}(\omega_x, \omega_y) = & [\hat{f}_t(\omega_x, \omega_y). \widehat{HH}(\omega_x, \omega_y)]. \widehat{\tilde{HH}}(\omega_x, \omega_y) \\ & + [\hat{f}_t(\omega_x, \omega_y). \widehat{GH}(\omega_x, \omega_y)]. \widehat{\tilde{GH}}(\omega_x, \omega_y) \\ & + [\hat{f}_t(\omega_x, \omega_y). \widehat{HG}(\omega_x, \omega_y)]. \widehat{\tilde{HG}}(\omega_x, \omega_y) \\ & + [\hat{f}_t(\omega_x, \omega_y). \widehat{GG}(\omega_x, \omega_y)]. \widehat{\tilde{GG}}(\omega_x, \omega_y) \end{aligned} \quad (8)$$

$$\begin{aligned} \hat{f}_{t+1}(\omega_x, \omega_y) = & \hat{f}_t(\omega_x, \omega_y). \{ [\widehat{HH}(\omega_x, \omega_y). \widehat{\tilde{HH}}(\omega_x, \omega_y)] \\ & + [\widehat{GH}(\omega_x, \omega_y). \widehat{\tilde{GH}}(\omega_x, \omega_y)] + [\widehat{HG}(\omega_x, \omega_y). \widehat{\tilde{HG}}(\omega_x, \omega_y)] \\ & + [\widehat{GG}(\omega_x, \omega_y). \widehat{\tilde{GG}}(\omega_x, \omega_y)] \} \end{aligned} \quad (9)$$

It is obvious from (9) that the perfect reconstruction (i.e., $f_{t+1}(x, y) = f_t(x, y)$) is guaranteed through inverse wavelet transform if,

$$\widehat{HH}. \widehat{\tilde{HH}} + \widehat{GH}. \widehat{\tilde{GH}} + \widehat{HG}. \widehat{\tilde{HG}} + \widehat{GG}. \widehat{\tilde{GG}} = 1 \quad (10)$$

which is usually the case for conventionally used orthogonal and biorthogonal filters (e.g., Haar, Daubechies, and Coiflet).

2.3 Wavelet Diffusion

In this section, we show that the standard discrete wavelet transform can be considered as having the same form as a single discrete step (4) of nonlinear diffusion. Equation (4) can be transformed into Fourier domain:

$$\begin{aligned} \hat{f}_{t+1}(\omega_x, \omega_y) = & \hat{f}_t(\omega_x, \omega_y) - \omega_x^2 \hat{f}_t(\omega_x, \omega_y) - \omega_y^2 \hat{f}_t(\omega_x, \omega_y) \\ & - i\omega_x \left[\frac{1}{2\pi} \hat{p}(\omega_x, \omega_y) \otimes i\omega_x \hat{f}_t(\omega_x, \omega_y) \right] \\ & - i\omega_y \left[\frac{1}{2\pi} \hat{p}(\omega_x, \omega_y) \otimes i\omega_y \hat{f}_t(\omega_x, \omega_y) \right] \end{aligned} \quad (11)$$

substituting $1 - \omega_x^2 - \omega_y^2 = \hat{A}_1 \cdot \hat{A}_2$, $i\omega_x = \hat{B}$, $-i\omega_x = (\hat{C} + \hat{D})$, $i\omega_y = \hat{E}$, $-i\omega_y = \hat{F}$, and $\frac{1}{2\pi} \hat{p}(\omega_x, \omega_y) = \hat{p}$, we obtain:

$$\begin{aligned}\hat{f}_{t+1}(\omega_x, \omega_y) &= \hat{A}_2 \cdot \hat{A}_1 \cdot \hat{f}_t(\omega_x, \omega_y) + \hat{C} \cdot [\hat{p} \otimes (\hat{B} \cdot \hat{f}_t(\omega_x, \omega_y))] \\ &\quad + \hat{D} \cdot [\hat{p} \otimes (\hat{B} \cdot \hat{f}_t(\omega_x, \omega_y))] + \hat{F} \cdot [\hat{p} \otimes (\hat{E} \cdot \hat{f}_t(\omega_x, \omega_y))]\end{aligned}\quad (12)$$

Taking the inverse Fourier transform of the above equation:

$$\begin{aligned}f_{t+1}(x, y) &= f_t(x, y) \otimes A_1 \otimes A_2 + [p(x, y) \cdot (f_t(x, y) \otimes B)] \otimes C \\ &\quad + [p(x, y) \cdot (f_t(x, y) \otimes B)] \otimes D \\ &\quad + [p(x, y) \cdot (f_t(x, y) \otimes E)] \otimes F\end{aligned}\quad (13)$$

interestingly, the perfect reconstruction condition (10) is also satisfied in this case by considering the values from the above substitutions,

$$\hat{A}_1 \cdot \hat{A}_2 + \hat{B} \cdot \hat{C} + \hat{B} \cdot \hat{D} + \hat{E} \cdot \hat{F} = 1 \quad (14)$$

Thus equations (12), (13), and (14) have equivalence with (8), (7), and (10), respectively. It is, therefore, deduced from this relationship of the nonlinear diffusion and DWT that wavelet based diffusion works in three steps:

- (i) wavelet decomposition into low-frequency subband (A_j) and high-frequency subbands (W_j^i);
- (ii) regularization of high-frequency coefficients (W_j^i) by multiplication with a function p ; i.e., $\tilde{W}_j^i = p_j * W_j^i$
- (iii) wavelet reconstruction from low-frequency subband (A_j) and regularized high-frequency subbands (\tilde{W}_j^i).

The regularization step multiplies the high-frequency wavelet coefficients by the regularization coefficient $p_j(|\eta_j(x, y)|) = 1 - g(|\eta_j(x, y)|)$ where $g(|\eta_j(x, y)|)$ is

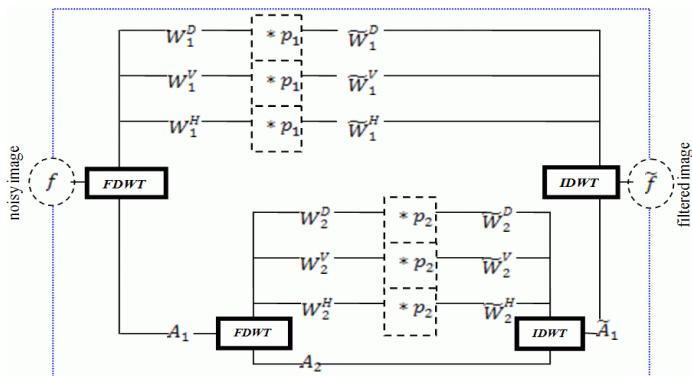


Fig. 1. Two-level wavelet diffusion

the aforementioned nonnegative monotonically decreasing diffusivity function and $|\eta_j(x, y)|$ is the wavelet domain edge estimate for pixel (x, y) at scale j . This process of discrete wavelet diffusion is illustrated in Fig. 1. It must be noted that the wavelet diffusion operation can be considered as an iterative wavelet shrinkage [2] process.

3 Experimental Results and Discussion

In this work, we used the undecimated discrete wavelet transform (UDWT) though DWT is also a possible option. The wavelet filter used was Haar and decomposition level was restricted to four, after initial experimentation with 1 to 5 decomposition levels. Further investigations on the use of decimated discrete wavelet transform, finding the appropriate number of decomposition levels, and the choice of optimal wavelet filter are left for future work. The diffusivity function g employed is the one proposed by Weickert et al. [9]:

$$g(|\eta_j(x, y)|) = \begin{cases} 1, & |\eta_j(x, y)| = 0 \\ 1 - \exp \left[\frac{-3.315}{\left(\frac{|\eta_j(x, y)|}{\lambda} \right)^4} \right], & |\eta_j(x, y)| > 0 \end{cases} \quad (15)$$

which is used to derive the regularization function p_j . The edge estimate $|\eta_j(x, y)|$ is approximated by the modulus of the high-frequency wavelet subbands at each scale j (i.e., each scale has its own edge estimate) and is computed by:

$$|\eta_j(x, y)| = \sqrt{\left(W_j^H(x, y) \right)^2 + \left(W_j^V(x, y) \right)^2 + \left(W_j^D(x, y) \right)^2} \quad (16)$$

A common value of edge threshold λ in (15) was used at all scales. The images were acquired from the “standard” test image collection of [10]. We used a multiscale approach for wavelet diffusion, in comparison to [4], by considering each approximation subband at a particular decomposition level as an independent image and high-frequency wavelet subbands of each level were diffused for noise removal. This lead to fewer iterations (about 5 on average) required for denoising in comparison to the single-scale diffusion of [4] which took nearly 20 or more iterations.

We experimented with different values for λ , the results of which are not shown (due to lack of space) where the noisy images were polluted with additive Gaussian noise of signal-to-noise-ratio (SNR) of 10dB and 15dB. Some visual results for denoising with the proposed wavelet diffusion, Wiener filtering and conventional nonlinear diffusion are presented in Fig. 2. Our results with varying the values of edge threshold λ are in line with the results of [4] and, in general, with the conventional nonlinear diffusion. It indicates that a higher value of λ takes lesser number of iterations to reach the best results but it is prone to oversmooth the image (rather than preserving the edges) in a few iterations. In contrast, a smaller value of λ takes a higher number of iterations to provide the best results but it usually avoids the oversmoothing effect. Moreover, the diffusion process should stop after a fixed

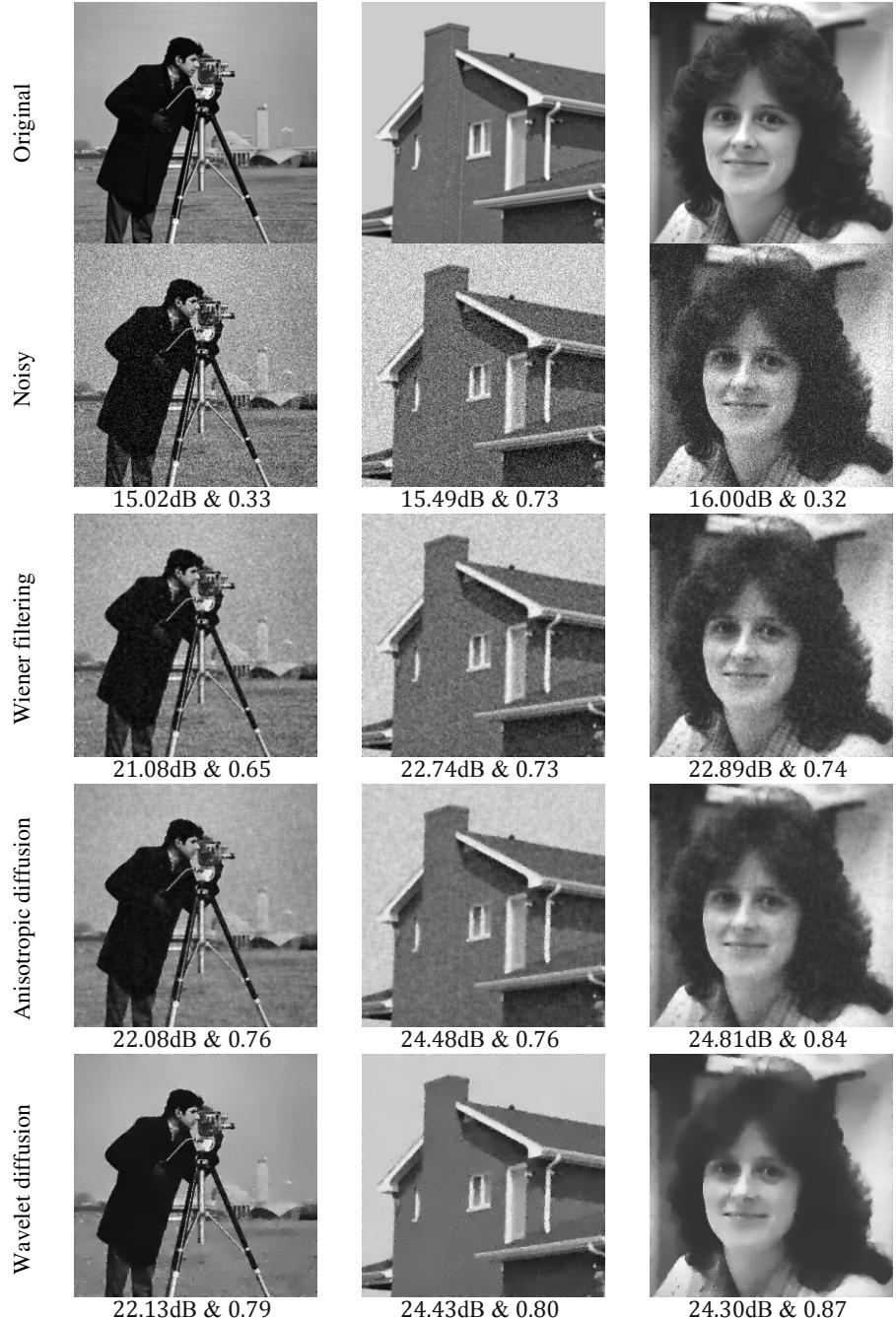


Fig. 2. Visual, SNR, and SSIM results (Anisotropic Diffusion and Wavelet Diffusion: $\lambda = 30$ and iterations = 3; Anisotropic Diffusion: Time-step = 0.5 and Weickert diffusivity function; Wiener filtering: 3x3 window)

number of iterations or it may actually degrade the image quality by oversmoothing effect. However, according to the authors' knowledge, there is currently no mechanism to determine the best number of diffusion iterations required (either in wavelet based or in conventional non-linear diffusion methods). In addition, we can deduce from these results that a higher value of λ provides good results when the amount of noise is high.

Fig. 2 shows the visual and quantitative results of the proposed wavelet diffusion and compares these with the results of Wiener filtering and conventional nonlinear diffusion in terms of visual quality, SNR, and structural similarity index measure (SSIM) [11]. Apparently, the SNR values of Wiener filtering and nonlinear diffusion for these images are close to the wavelet diffusion results but the visual quality and SSIM values (in the range of 0 to 1) indicate that the results with the proposed method have superior edge preservation, region homogeneity, and similarity to the original image. It was found in [11] that SNR may give flawed measure of image quality as it is not sensitive to the structure (e.g., edges) in an image. It is evident that wavelet diffusion method either provides comparable or favourable results than both Wiener filtering and Weickert nonlinear diffusion. We experimented with a range of parameter values and chose those values, specified in Fig. 2, that produced optimal results with the given set of images.

4 Conclusions

In this paper, we have proposed a generalization of the wavelet diffusion for use with the discrete wavelet transform. We demonstrated that a single discrete step of nonlinear diffusion can be considered as equivalent to the DWT decomposition, high-frequency subbands regularization, and reconstruction. We showed also that the wavelet diffusion can operate in multiscale. Image denoising results of the proposed method compared with Wiener filtering and conventional nonlinear diffusion showed that the technique is promising. Moreover, the method smoothes the noise while preserving the important structures such as edges. Further investigations could be done in the use of DWT instead of UDWT, the number of wavelet decomposition levels, the wavelet bases, number of diffusion iterations, comparison with MZ-DWT based diffusion, and using varying threshold value for different scale. This may require looking for an automatic method of threshold computation.

References

- [1] Perona, P., Malik, J.: Scale Space and Edge Detection Using Anisotropic Diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12, 629–639 (1990)
- [2] Donoho, D.L., Johnstone, I.M.: Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika* 81, 425–455 (1994)
- [3] Coifman, R.R., Donoho, D.: Translation-Invariant Denoising. In: Antoine, A., Oppenheim, G. (eds.) *Wavelets in Statistics*, pp. 125–150. Springer, New York (1995)
- [4] Shih, A.C.-C., Liao, H.-Y.M., Lu, C.-S.: A New Iterated Two-Band Diffusion Equation: Theory and Its Applications. *IEEE Transactions on Image Processing* (2003)

- [5] Yue, Y., Croitoru, M.M., Bidani, A., Zwischenberger, J.B., Jr., J.W.C.: Nonlinear Multiscale Wavelet Diffusion for Speckle Suppression and Edge Enhancement in Ultrasound Images. *IEEE Transactions on Medical Imaging* 25, 297–311 (2006)
- [6] Mallat, S., Zhong, S.: Characterization of Signals from Multiscale Edges. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14, 710–732 (1992)
- [7] Mrazek, P., Weickert, J., Steidl, G.: Diffusion-Inspired Shrinkage Functions and Stability Results for Wavelet Denoising. *International Journal of Computer Vision* 64, 171–186 (2005)
- [8] Mallat, S.: *A Wavelet Tour of Signal Processing*. Academic Press, MA (1998)
- [9] Weickert, J.: *Anisotropic Diffusion in Image Processing*. Stuttgart, Germany: Teubner (1998)
- [10] <http://www.imageprocessingplace.com>
- [11] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* 13, 600–612 (2004)

Dynamic Color Texture Modeling and Color Video Decomposition Using Bounded Variation and Oscillatory Functions

Mathieu Lugez^{1,3}, Michel Ménard^{1,3}, and Abdallah El-Hamidi^{2,3}

¹ L3i - Université de La Rochelle

² MIA - Université de La Rochelle

³ Fédération PRIDES: Pôle régional de Recherche en Images, Données et Systèmes

Abstract. Dynamic, or temporal, texture is a spatially repetitive, time-varying visual pattern that forms an image sequence with a certain temporal stationarity. Important tasks are thus the detection, segmentation and perceptual characterization of Dynamic Texture (DT). Following recent work, color image decomposition appears as a good way to reach these different aims, however, to our best knowledge, no proposed model is currently able to deal with temporal aspect, inherent to color image sequences.

The major contribution of this paper is to adapt static decomposition model to time aspect in order to deal with videos and color image sequences. In this paper we propose an extended decomposition model which splits a video into two components, a first one containing geometrical information, the structure of the sequence and a second one dynamic color texture and noise. Examples for color video decomposition and characterization of real dynamic present in texture component will be presented.

1 Introduction

1.1 Motivation

A new issue in texture analysis is its extension to the temporal domain, a field known as Dynamic Texture (DT) analysis. In DT, the notion of self-similarity central to conventional image textures is extended to the spatiotemporal domain. DT are typically result from processes such as of water flows, smoke, fire, a flag blowing in the wind, a moving escalator, or a walking crowd. Important tasks are thus the detection, segmentation and perceptual characterization of dynamic textures. The ultimate goal is to be able to support video queries based on the recognition of the actual natural and artificial dynamic texture processes.

Following recent work, color image decomposition into geometrical and texture components appears as a good way to reach this aim in extracting meaningful information, i.e texture component, independently of geometrical information.

In this way, we propose to extend spatial color decomposition model to spatiotemporal domain, and attempt to highlight time influence present in video to characterize dynamic texture. To the best of our knowledge, no such time adaptation is currently available.

1.2 Overview of the Paper

The aim of this work is to extend a model, which decompose color image into two components, a first one containing geometrical structure U and a second one V , holding textural information and noise. So, we aim to deal with color image sequences in extending to time existing reliable model. Moreover, through decomposition of texture component, we attempt to determinate time impact, showing up dynamicity present in sequences, which will be suit for future work on dynamic texture.

In the first place of this paper we introduce the extended minimization functional problem and the associate discrete framework in which we place ourself and which is an appropriate one in image sequence processing. In a second part we present the extended time decomposition model and subsequently its color implementation and choice of parameters. Then, in the last part, we show some significant results and attempt to present time weight in dynamic texture extraction through color image sequence.

2 Time Extension of Decomposition Model

Decomposing an image into meaningful component appears as one of major aims in recent development in image processing. The first goal was image restoration and demising; but following the ideas of Yves Meyer, in total variation minimization framework of L. Rudin, S. Osher and E. Fatemi, image decomposition into geometrical and oscillatory (i.e texture) component appears as usefull and very interesting way in computer vision and image analysis. There is a very large literature and recent advances on image decomposition models, image regularization and texture extraction and modelling. So, we only cite, among many others, most recent works which we appear like most relevant and usefull paper. In this way, reader can refer to the work of Stark et al. [1], Aujol et al. [2], [3], Aujol and Chambolle [4], Aujol and Ha Kang [5] and Vese and Osher [6], [7], [8] to cover the most recent and relevant advances.

2.1 Decomposition Model and Functional Spaces

In order to decompose image sequences in suitable components we propose to extend the Osher-Vese [8] color decomposition model. Their approach derived from Meyer decomposition model [9]. Its rely on space of functions of bounded variation, BV and Banach space, G . Authors propose to minimize the following discretized functional:

$$\inf_{(u,v) \in BV \times G / f = u + v} \left\{ F(u, v) = \int |\nabla u| + \lambda \|v\|_G \right\} \quad (1)$$

BV is a good space to model functions with discontinuities along lines and curves, that can represent and preserve edges in an image. For the reader convenience we recall here its definition [10]:

Definition 1. $BV(\Omega)$ is the subspace functions $u \in L^1(\Omega)$ such that the following quantity, called the total variation of u , is finite:

$$J(u) = \sup \left\{ \int_{\Omega} u(x) \operatorname{div}(\xi(x)) dx \right\} \quad (2)$$

such that $\xi \in C_c^1(\Omega, \mathbb{R}^2)$, $\|\xi\|_{L^\infty(\Omega)} \leq 1$

In order to improve representation of oscillating patterns, Meyer in [9] has introduced the Banach space, G , which modeling signals with large oscillations, in particular textures and noise. A function belonging to G may have large oscillations and nevertheless have a small norm. Thus the norm on G is well-adapted to capture the oscillations of a function in an energy minimization method. We recall here the definition of G :

Definition 2. G is the Banach space composed of the distributions f which can be written $f = \partial_1 g_1 + \partial_2 g_2 = \operatorname{div}(g)$ with g_1 and g_2 in $L^\infty(\Omega)$. On G , the following norm is associate:

$$\|v\|_G = \inf \left\{ \|g\|_{L^\infty(\Omega, \mathbb{R}^2)} / v = \operatorname{div}(g), g = (g_1, g_2), |g(x)| = \sqrt{(|g_1|^2 + |g_2|^2)(x)} \right\} \quad (3)$$

2.2 Spatiotemporal Structure and Discretization

To take into account the spatiotemporal structure, we consider a video as an 3-D image [11], i.e a volume, so that we can apply 2-D image algorithms extended to the 3-D case. We assume that we have a given image sequence $f \in L^2(\Omega)$, where Ω is an open and bounded domain on \mathbb{R}^3 , with Lipschitz boundary. In order to recover u and v from f , we propose:

- An extended discrete version of gradient vector $|\nabla u|$ given by:

$$(\nabla u)_{i,j,k} = ((\nabla u)_{i,j,k}^1, (\nabla u)_{i,j,k}^2, (\nabla u)_{i,j,k}^3) \quad (4)$$

$$(\nabla u)_{i,j,k}^1 = \begin{cases} u_{i+1,j,k} - u_{i,j,k} & \text{if } i < N \\ 0 & \text{if } i = N \end{cases}$$

$$(\nabla u)_{i,j,k}^2 = \begin{cases} u_{i,j,k} - u_{i,j+1,k} & \text{if } j < N \\ 0 & \text{if } j = N \end{cases}$$

$$(\nabla u)_{i,j,k}^3 = \begin{cases} u_{i,j,k} - u_{i,j,k-1} & \text{if } k < N \\ 0 & \text{if } k = N \end{cases}$$

- An extended discrete total variation definition:

$$J(u) = (\nabla u)_{i,j,k}^1 + (\nabla u)_{i,j,k}^2 + \alpha (\nabla u)_{i,j,k}^3 \quad (5)$$

we introduce the α constant to maintain homogeneity between space and time component. It's mainly for numerical implementation, to avoid discretization problem due to quantization step, which be different along space and time dimension. In practice, we often set it to one, but user can adapt it to less, more or in function of frame per second, or quickness of movement present in sequence, to ensure most reliability and homogeneity.

- An adapted definition of G , inspired by [8] for the vector case, extended to the third dimension:

Definition 3. Let G denote the Banach space consisting of all generalized vector-valued functions $\vec{v}(x, y, t) = (v_R(x, y, t), v_G(x, y, t), v_B(x, y, t))$ which can be written as:

$$\vec{v}(x, y, t) = (\operatorname{div} \vec{g}_R, \operatorname{div} \vec{g}_G, \operatorname{div} \vec{g}_B) \quad (6)$$

$$g_{1,c}, g_{2,c}, g_{3,c} \in L^\infty(\mathbb{R}^3), c = R, G, B,$$

induced by the norm $\|v\|_*$ defined as the lower bound of all L^∞ norms of functions $|\vec{g}|$, where $|\vec{g}| = \sqrt{|\vec{g}_R|^2 + |\vec{g}_G|^2 + |\vec{g}_B|^2} = \sqrt{\sum_{c=R,G,B} ((g_{1,c})^2 + (g_{2,c})^2 + (g_{3,c})^2)}$, and where the infinitum is computed over all decompositions (6) of \vec{v} .

3 Description of the Extended Decomposition Model

For the reader convenience we only use, by now, useful indices in our formulas to avoid overloaded notation, so we will use c indices to denote color channel ($c \in R, G, B$) and 1, 2 and 3 exponents to represent the dimension to which objects belong.

3.1 Description of the Model

We propose the following minimization problem inspired by (6) and [8], for each color channel:

$$\inf_{u, g_1, g_2, g_3} \left\{ G_p(u, g_1, g_2, g_3) = \int |\nabla u| + \lambda \int |f - u - \partial_x g_1 - \partial_y g_2 - \partial_t g_3|^2 dx dy dt \right. \\ \left. + \mu \left[\int (\sqrt{g_1^2 + g_2^2 + g_3^2})^p dx dy dt \right]^{\frac{1}{p}} \right\} \quad (7)$$

where λ, μ are tuning parameters (see next section for more useful details). Let reintroduce that $\vec{u} = (u_R, u_G, u_B)$, and $\vec{g}_i = (\vec{g}_{i,R}, \vec{g}_{i,G}, \vec{g}_{i,B})$, $i \in \{1, 2, 3\}$.

Formally minimizing the above energy equation with respect to u, g_1, g_2, g_3 , yields the following Euler-Lagrange equation for each color channel:

$$u = f - \partial_x g_1 - \partial_y g_2 - \partial_t g_3 + \frac{1}{2\lambda} \operatorname{div} \left(\frac{\nabla u}{|\nabla u|} \right) \quad (8)$$

$$\mu(\|\sqrt{g_1^2 + g_2^2 + g_3^2}\|)^{1-p} (\sqrt{g_1^2 + g_2^2 + g_3^2})^{p-2} g_1 = 2\lambda \left[\frac{\partial}{\partial_x} (u - f) + \partial_{xx}^2 g_1 + \partial_{xy}^2 g_2 + \partial_{xt}^2 g_3 \right] \quad (9)$$

$$\mu(\|\sqrt{g_1^2 + g_2^2 + g_3^2}\|)^{1-p} (\sqrt{g_1^2 + g_2^2 + g_3^2})^{p-2} g_2 = 2\lambda \left[\frac{\partial}{\partial_y} (u - f) + \partial_{yx}^2 g_1 + \partial_{yy}^2 g_2 + \partial_{yt}^2 g_3 \right] \quad (10)$$

$$\mu(\|\sqrt{g_1^2 + g_2^2 + g_3^2}\|)^{1-p} (\sqrt{g_1^2 + g_2^2 + g_3^2})^{p-2} g_3 = 2\lambda \left[\frac{\partial}{\partial_t} (u - f) + \partial_{xt}^2 g_1 + \partial_{yt}^2 g_2 + \partial_{tt}^2 g_3 \right] \quad (11)$$

To simplify the presentation, let us introduce the notation:

$$H(g_1, g_2, g_3) = (\|\sqrt{g_1^2 + g_2^2 + g_3^2}\|)^{1-p} (\sqrt{g_1^2 + g_2^2 + g_3^2})^{p-2}$$

3.2 Numerical Implementation

For the reader convenience, we present here a numerical scheme to solve our minimization problem. So to solve equations (8) to (11) we use a semi-implicit finite differences scheme and an iterative algorithm, based on fixed point iteration [6].

We got, in order to compute gradient:

$$|\nabla u| \doteq |\nabla u|_{i,j,k} = \sqrt{\left(\frac{u_{i+1} - u_{i-1}}{2h}\right)^2 + \left(\frac{u_{j+1} - u_{j-1}}{2h}\right)^2 + \left(\frac{u_{k+1} - u_{k-1}}{2h}\right)^2} \quad (12)$$

so, belong first dimension (x) we got:

$$\begin{aligned} |\nabla u|_{i+\frac{1}{2}, j, k} &= \sqrt{\left(\frac{u_{i+1}-u}{h}\right)^2 + \left(\frac{u_{j+1}-u_{j-1}}{2h}\right)^2 + \left(\frac{u_{k+1}-u_{k-1}}{2h}\right)^2} \\ |\nabla u|_{i-\frac{1}{2}, j, k} &= \sqrt{\left(\frac{u-u_{i-1}}{h}\right)^2 + \left(\frac{u_{i-1, j+1}-u_{i-1, j-1}}{2h}\right)^2 + \left(\frac{u_{i-1, k+1}-u_{i-1, k-1}}{2h}\right)^2} \end{aligned} \quad (13)$$

in the same way, we compute gradient belong y and t dimension.

We are now in position to introduced complete numerical scheme to solving Euler-Lagrange equations:

$$\begin{aligned} u^{n+1} &= \frac{1}{1 + \frac{1}{3\lambda h^2} \left(\frac{1}{|\nabla|_{i+\frac{1}{2}}} + \frac{1}{|\nabla|_{i-\frac{1}{2}}} + \frac{1}{|\nabla|_{j+\frac{1}{2}}} + \frac{1}{|\nabla|_{j-\frac{1}{2}}} + \frac{1}{|\nabla|_{k+\frac{1}{2}}} + \frac{1}{|\nabla|_{k-\frac{1}{2}}} \right)} \\ &\quad * \left[f - \partial_x g_1 - \partial_y g_2 - \partial_t g_3 + \frac{1}{3\lambda h^2} \left(\frac{u_{i+1}^n}{|\nabla|_{i+\frac{1}{2}}} + \frac{u_{i-1}^n}{|\nabla|_{i-\frac{1}{2}}} + \frac{u_{j+1}^n}{|\nabla|_{j+\frac{1}{2}}} + \frac{u_{j-1}^n}{|\nabla|_{j-\frac{1}{2}}} + \frac{u_{k+1}^n}{|\nabla|_{k+\frac{1}{2}}} + \frac{u_{k-1}^n}{|\nabla|_{k-\frac{1}{2}}} \right) \right] \end{aligned} \quad (14)$$

$$g_1^{n+1} = \frac{3\lambda}{\mu H(g_1^n, g_2^n, g_3^n) + \frac{6\lambda}{h^2}} \left[\partial_x u^n - \partial_x f + \frac{g_{1,i+1}^n + g_{1,i-1}^n}{h^2} + \partial_{xy}^2 g_2^n + \partial_{xt}^2 g_3^n \right] \quad (16)$$

$$g_2^{n+1} = \frac{3\lambda}{\mu H(g_1^n, g_2^n, g_3^n) + \frac{6\lambda}{h^2}} \left[\partial_y u^n - \partial_y f + \partial_{xy}^2 g_1^n + \frac{g_{2,j+1}^n + g_{2,j-1}^n}{h^2} + \partial_{yt}^2 g_3^n \right] \quad (17)$$

$$g_3^{n+1} = \frac{3\lambda}{\mu H(g_1^n, g_2^n, g_3^n) + \frac{6\lambda}{h^2}} \left[\partial_t u^n - \partial_t f + \partial_{xt}^2 g_1^n + \partial_{yt}^2 g_2^n + \frac{g_{3,k+1}^n + g_{3,k-1}^n}{h^2} \right] \quad (18)$$

(19)

where h denote the discrete step space, so in practice we set $h = 1$ in our algorithm, moreover in furtherance of time computing, we always use the most recent computed values. For the initialization of parameters we set: $u^0 = f$, $g_1^0 = \frac{-1}{3\lambda} \frac{\partial_x f}{\sqrt{\partial_x f^2 + \partial_y f^2 + \partial_t f^2}}$, $g_2^0 = \frac{-1}{3\lambda} \frac{\partial_y f}{\sqrt{\partial_x f^2 + \partial_y f^2 + \partial_t f^2}}$ and $g_3^0 = \frac{-1}{3\lambda} \frac{\partial_t f}{\sqrt{\partial_x f^2 + \partial_y f^2 + \partial_t f^2}}$

3.3 Choice of Parameters

To obtain weak regularization, we use parameters as: $\lambda = 0.1$, $\mu = 0.1$ and we compute about 20 iterations. For strong regularization we got: $\lambda = 0.005$, $\mu = 0.005$ iterated 50 times. So, for classic decomposition we use parameters as: $\lambda = 0.01$, $\mu = 0.01$ iterated 20 times. For all results present in the next section we computed our decomposition on 8 images treated simultaneously as one block, but our algorithm is able to deal with bigger image block to catch lower frequencies and wider range of time oscillating patterns.

4 Numerical Results and Time Component

All images and results are compute from DynTex, the dynamic texture database [12] which provide a large and diverse database of high-quality dynamic textures. Dyntex sequences come from natural scene presenting a wide variety of moving process as flowing water, leaves blowing in wind, walking crowd... Such diversity grants user to identify and emphasize a lot of aspects in testing purpose.

For more details, demonstration sequence, wider range of results and for a presentation of similar color image sequences decomposition, relying on different approaches in the same framework [13], please consult this URL: <http://perso.univ-lr.fr/mlugiez>.

4.1 Static vs. Dynamic Decomposition

In order to prove that our dynamic decomposition method show more significant result than a static decomposition, we present a comparison between two

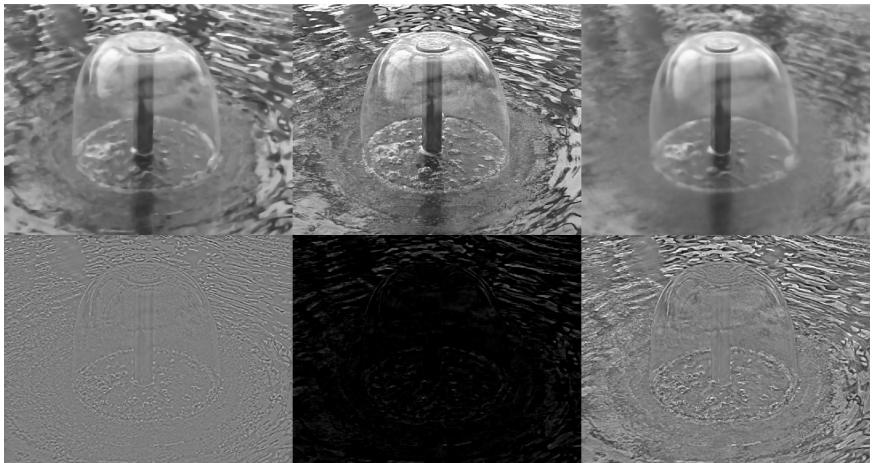


Fig. 1. From left to right, the geometrical component, U , in classic color decomposition (top) and its texture and noise component V (bottom). The original image (center and top), the difference between static and dynamic V component (center and bottom). Then our new dynamic decomposition components (right).

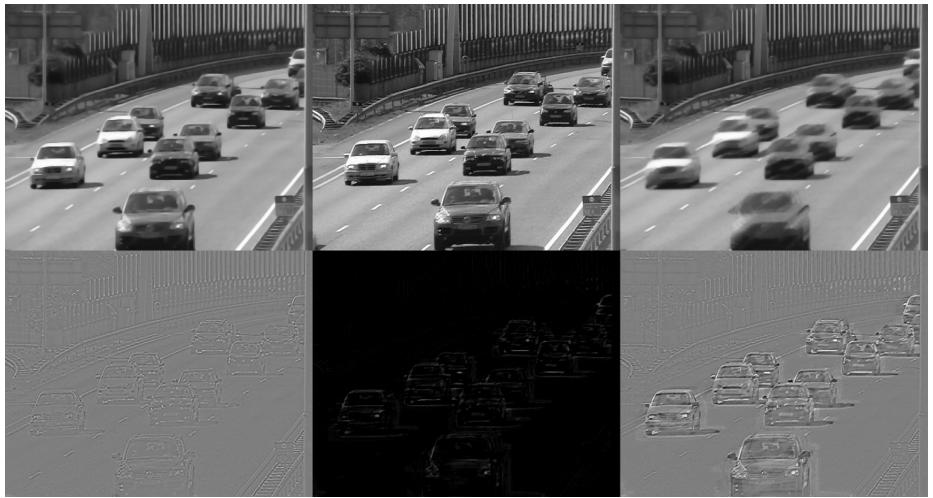


Fig. 2. From left to right, the geometrical component, U , in classic color decomposition (top) and its texture and noise component V (bottom). The original image (center and top), the difference between static and dynamic V component (center and bottom). Then our new dynamic decomposition components (right).

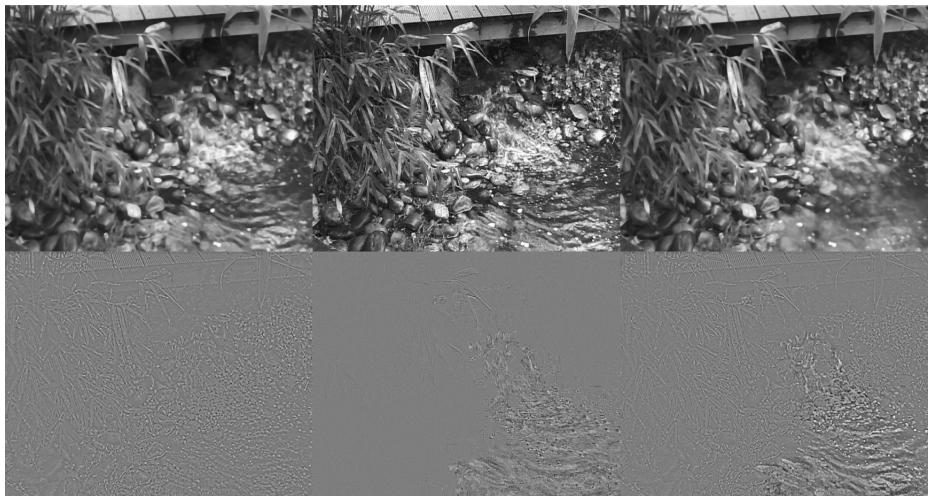


Fig. 3. From left to right, the geometrical component, U , in classic color decomposition (top) and its texture and noise component V (bottom). The original image (center and top), the time influence g_3 part of V component (center and bottom). Then our new dynamic decomposition components (right).

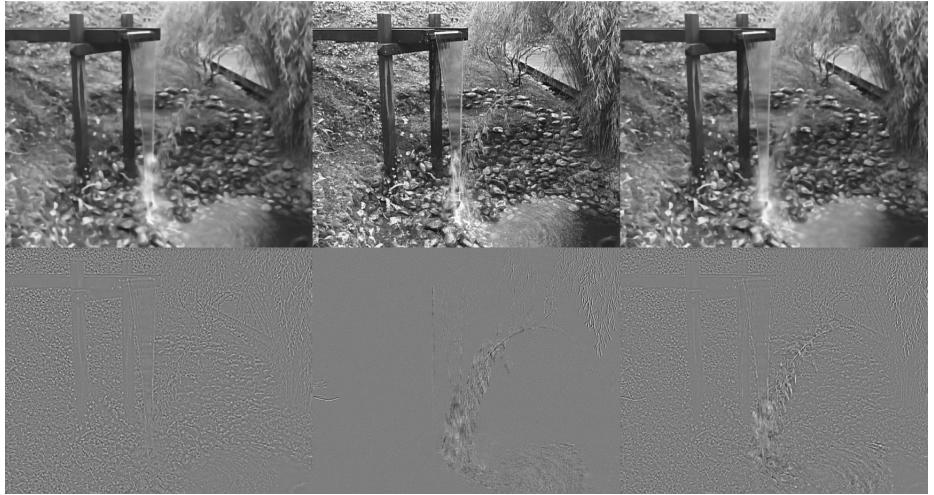


Fig. 4. From left to right, the geometrical component, U , in classic color decomposition (top) and its texture and noise component V (bottom). The original image (center and top), the time influence g_3 part of V component (center and bottom). Then our new dynamic decomposition components (right).

methods (static and dynamic decomposition are both computed with same classic parameters). We can easily see that time impact in result, water in Fig.1 is well regularized and fluid aspect is well represent in the V component. In Fig. 2 we can clearly see the reenforcement of moving cars texture without that static part and objects present in sequence are taken into account. Moreover the regularization is more robust to illumination and movement constrain.

Moreover, if user tunes parameters to obtain stronger regularization, our algorithm is able to catch wider waves in spatiotemporal texture component: see the circumference of fountain in Fig. 4, more regularized (in U component) than wider waves. It's a matter of deep in spatiotemporal texture extraction, which our algorithm is able to deal with.

4.2 Time Impact in Our Decomposition

We present, in Fig. 3, a part of a decomposed sequence of flowing water under wood bridge. We can see the static aspect of U component, regularized in space and in time, seems to be freezed, although texture component, V , present a real dynamic, strengthened by time influence. Only moving things or objects presenting dynamicity are taken in account in the third part of V component (i.e g_3). In Fig. 4 we can distinctly see time influence, reed's branch oscillating under water flow is clearly highlight, waves present in basin's fountain are well regularized in U component, water dynamicity is totally catch as texture. We can clearly see that is due to the third part of texture component, movement information is well captured in this component. In this way we obtain the dynamicity present

in video through oscillations along time dimension. These results will be useful for future work on dynamic texture characterization.

References

1. Starck, J.L., Elad, M., Donoho, D.L.: Image decomposition via the combination of sparse representations and a variational approach. *IEEE Trans. Image Processing* 14(10), 1570–1582 (2005)
2. Aujol, J.F., Aubert, G., Feraud, L.B., Chambolle, A.: Image decomposition into a bounded variation component and an oscillating component. *Journal of Mathematical Imaging and Vision* 22(1), 71–88 (2005)
3. Aujol, J.F., Gilboa, G., Chan, T., Osher, S.: Structure-texture image decomposition - modeling, algorithms, and parameter selection. *International Journal of Computer Vision* 67(1), 111–136 (2006)
4. Aujol, J.F., Chambolle, A.: Dual norms and image decomposition models. *International Journal of Computer Vision* 63(1), 85–104 (2005)
5. Aujol, J.F., Kang, S.H.: Color image decomposition and restoration. *J. Visual Communication and Image Representation* 17(4), 916–928 (2006)
6. Vese, L.A., Osher, S.J.: Modeling textures with total variation minimization and oscillating patterns in image processing. *J. Sci. Computing* 19, 553–572 (2003)
7. Vese, L.A., Osher, S.J.: Image denoising and decomposition with total variation minimization and oscillatory functions. *Journal of Mathematical Imaging and Vision* 20(1-2), 7–18 (2004)
8. Vese, L.A., Osher, S.: Color texture modeling and color image decomposition in a variational-PDE approach. In: SYNASC, pp. 103–110. IEEE Computer Society, Los Alamitos (2006)
9. Meyer, Y.: Oscillating patterns in image processing and nonlinear evolution equations. University Lecture Series, vol. 22, American Mathematical Society, Providence, RI, The fifteenth Dean Jacqueline B. Lewis memorial lectures (2001)
10. Ambrosio, L., Fusco, N., Pallara, D.: Functions of bounded variation and free discontinuity problems. Oxford Mathematical Monographs. Oxford Mathematical Monographs. Clarendon Press, Oxford (2000)
11. Aubert, G., Kornprobst, P.: Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations, 2nd edn. Applied Mathematical Sciences, vol. 147. Springer, Heidelberg (2006)
12. Péteri, R., Huskies, M., Fazekas., S.: Dyntex: A comprehensive database of dynamic textures (2005), <http://www.cwi.nl/projects/dyntex/>
13. Lugiez, M., Ménard, M., El-Hamidi, A.: Spatiotemporal extension of color decomposition model and dynamic color structure-texture extraction (2008)



RETRACTED CHAPTER: Reinstating Floyd-Steinberg: Improved Metrics for Quality Assessment of Error Diffusion Algorithms

Sam Hocevar¹ and Gary Niger²

¹ Laboratoire d'Imagerie Bureautique et de Conception Artistique
14 rue de Plaisance, Paris, France

² 143 Rolloffle Avenue, Tarzana, California 91356
sam@hocevar.net, gary_niger@gnaa.us

Abstract. In this contribution we introduce a little-known property of error diffusion halftoning algorithms which we call *error diffusion displacement*. By accounting for the inherent sub-pixel displacement caused by the error propagation, we correct an important flaw in most metrics used to assess the quality of resulting halftones. We find these metrics to usually highly underestimate the quality of error diffusion in comparison to more modern algorithms such as direct binary search. Using empirical observation, we give a method for creating computationally efficient, image-independent, model-based metrics for this quality assessment. Finally, we use the properties of error diffusion displacement to justify Floyd and Steinberg's well-known choice of algorithm coefficients.

Keywords: halftoning, error diffusion, image quality, human visual system, color quantization.

1 Introduction

Image dithering is the process of reducing continuous-tone images to images with a limited number of available colours. Applications vary tremendously, from laser and ink-jet printing to display on small devices such as cellphones, or even the design of banknotes.

Countless methods have been published for the last 40 years that try to best address the problem of colour reduction. Comparing two algorithms in terms of speed or memory usage is often straightforward, but how exactly a halftoning algorithm performs quality-wise is a far more complex issue, as it highly depends on the display device and the inner workings of the human eye.

Though this document focuses on the particular case of bilevel halftoning, most of our results can be directly adapted to the more generic problem of colour reduction.

2 Halftoning Algorithms

The most ancient halftoning method is probably classical screening. This highly parallelisable algorithm consists in tiling a dither matrix over the image and

The original version of this chapter was retracted: The retraction note to this chapter is available at https://doi.org/10.1007/978-3-540-69905-7_71

using its elements as threshold values. Classical screening is known for its structural artifacts such as the cross-hatch patterns caused by Bayer ordered dither matrices [1]. However, modern techniques such as the void-and-cluster method [2], [3] allow to generate screens yielding visually pleasing results.

Error diffusion dithering, introduced in 1976 by Floyd and Steinberg [4], tries to compensate for the thresholding error through the use of feedback. Typically applied in raster scan order, it uses an error diffusion matrix such as the following one, where x denotes the pixel being processed:

$$\frac{1}{16} \begin{vmatrix} -x & 7 \\ 3 & 5 & 1 \end{vmatrix}$$

Though efforts have been made to make error diffusion parallelizable [5], it is generally considered more computationally expensive than screening, but carefully chosen coefficients yield good visual results [6].

Model-based halftoning is the third important algorithm category. It relies on a model of the human visual system (HVS) and attempts to minimise an error value based on that model. One such algorithm is direct binary search (DBS) [10], also referred to as least-squares model-based halftoning (LSMB) [16].

HVS models are usually low-pass filters. Nasanen [9], Analoui and Allebach [10] found that using Gaussian models gave visually pleasing results, an observation confirmed by independent visual perception studies [11].

DBS yields halftones of impressive quality. However, despite efforts to make it more efficient [12], it suffers from its large computational requirements and error diffusion remains a more widely used technique.

3 Error Diffusion Displacement

Most error diffusion implementations parse the image in raster scan order. Boustrophedonic (serpentine) scanning has been shown to cause fewer visual artifacts [7], but other, more complex processing paths such as Hilbert curves [8] are seldom used as they do not improve the image quality significantly.

Intuitively, as the error is always propagated to the bottom-left or bottom-right of each pixel (Fig. 1), one may expect the resulting image to be slightly translated. This expectation is confirmed visually when rapidly switching between an error diffused image and the corresponding DBS halftone.

This small translation is visually innocuous but we found that it means a lot in terms of error computation. A common way to compute the error between an image $h_{i,j}$ and the corresponding binary halftone $b_{i,j}$ is to compute the mean square error between modified versions of the images, in the form:

$$E(h, b) = \frac{(\|v * h_{i,j} - v * b_{i,j}\|_2)^2}{wh} \quad (1)$$

where w and h are the image dimensions, $*$ denotes the convolution and v is a model for the human visual system.

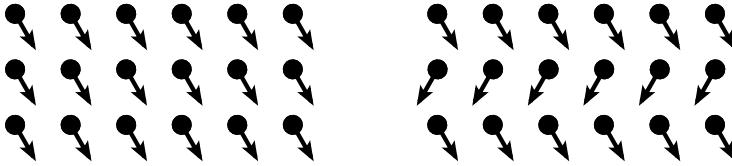


Fig. 1. Floyd-Steinberg error diffusion direction in raster scan (left) and serpentine scan (right)

To compensate for the slight translation observed in the halftone, we use the following error metric instead:

$$E_{dx,dy}(h, b) = \frac{(||v * h_{i,j} - v * t_{dx,dy} * b_{i,j}||_2)^2}{wh} \quad (2)$$

where $t_{dx,dy}$ is an operator which translates the image along the (dx, dy) vector. By design, $E_{0,0} = E$.

A simple example can be given using a Gaussian IIS model:

$$v(x, y) = e^{\frac{x^2+y^2}{2\sigma^2}} \quad (3)$$

Finding the second filter is then straightforward:

$$(v * t_{dx,dy})(x, y) = e^{\frac{(x-dx)^2+(y-dy)^2}{2\sigma^2}} \quad (4)$$

Experiments show that for a given image and a given corresponding halftone, $E_{dx,dy}$ has a local minimum almost always away from $(dx, dy) = (0, 0)$ (Fig. 2). Let E be an error metric where this remains true. We call the local minimum E_{min} :

$$E_{min}(h, b) = \min_{dx,dy} E_{dx,dy}(h, b) \quad (5)$$

For instance, a Floyd-Steinberg dither of *Lena* with $\sigma = 1.2$ yields a per-pixel mean square error of 3.67×10^{-4} . However, when taking the displacement into account, the error becomes 3.06×10^{-4} for $(dx, dy) = (0.165, 0.293)$. The new, corrected error is significantly smaller, with the exact same input and output images.

Experiments show that the corrected error is always noticeably smaller except in the case of images that are already mostly pure black and white. The experiment was performed on a database of 10,000 images from common computer vision sets and from the image board *4chan*, providing a representative sampling of the photographs, digital art and business graphics widely exchanged on the Internet nowadays [13].

In addition to the classical Floyd-Steinberg and Jarvis-Judice-Ninke kernels, we tested two serpentine error diffusion algorithms: Ostromoukhov's simple error diffusion [15], which uses a variable coefficient kernel, and Wong and Allebach's optimum error diffusion kernel [14]:

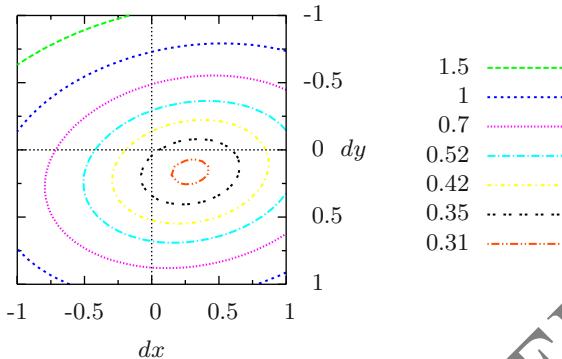


Fig. 2. Mean square error for the *Lena* image. v is a simple 11×11 Gaussian convolution kernel with $\sigma = 1.2$ and (dx, dy) vary in $[-1, 1] \times [-1, 1]$

	$E \times 10^4$	$E_{min} \times 10^4$
raster Floyd-Steinberg	3.7902	3.1914
raster Ja-Ju-Ni	9.7013	6.6349
Ostromoukhov	4.6812	4.4783
optimum kernel	5.5209	6.5772

We clearly see that usual metrics underestimate the quality of error-diffused halftones, especially in raster scan. Algorithms such as direct binary search, on the other hand, do not suffer from this bias since they are designed to minimise the very error induced by the HVS model.

4 An Image-independent Corrected Quality Metric for Error-Diffused Halftones

We have seen that for a given image, $E_{min}(h, b)$ is a better and fairer visual error measurement than $E(h, b)$. However, its major drawback is that it is highly computationally expensive: for each image, the new (dx, dy) values need to be calculated to minimise the error value.

Fortunately, we found that for a given raster or serpentine scan error diffusion algorithm, there was often very little variation in the optimal (dx, dy) values (Fig. 3 and 4).

For each algorithm, we choose the (dx, dy) values at the histogram peak and we refer to them as the *algorithm's displacement*, as opposed to the *image's displacement* for a given algorithm. We call $E_{fast}(h, b)$ the error computed at (dx, dy) . As E_{fast} does not depend on the image, it is a lot faster to compute than E_{min} , and as it is statistically closer to E_{min} , we can expect it to be a better error estimation than E :

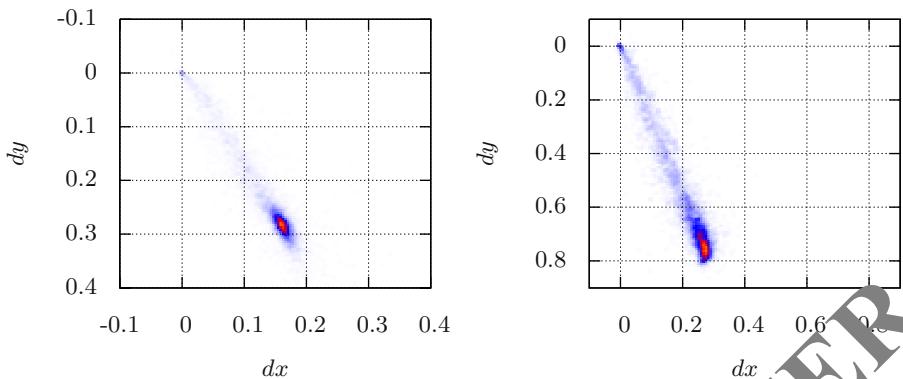


Fig. 3. Error diffusion displacement histograms for the raster Floyd-Steinberg (left) and raster Jarvis, Judis and Ninke (right) algorithms applied to a corpus of 10,000 images

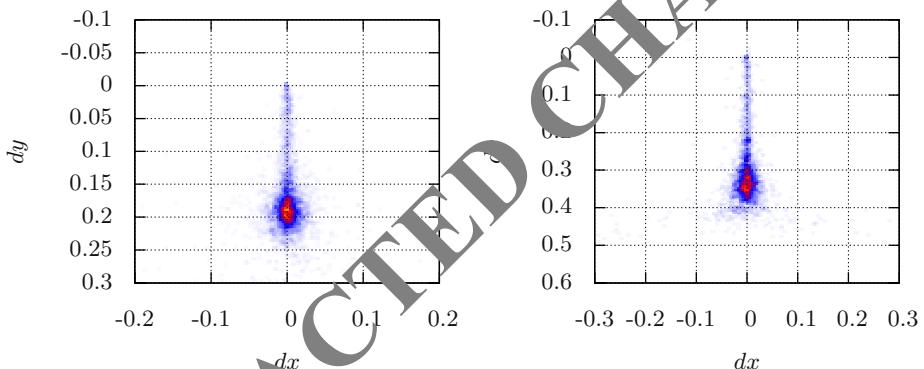


Fig. 4. Error diffusion displacement histograms for the Ostromoukhov (left) and optimum kernel (right) algorithms applied to a corpus of 10,000 images

	$E \times 10^4$	$E_{min} \times 10^4$	dx	dy	$E_{fast} \times 10^4$
raster Floyd-Steinberg	3.7902	3.1914	0.16	0.28	3.3447
raster JaJu-Ni	9.7013	6.6349	0.26	0.76	7.5891
Ostromoukhov	4.6892	4.4783	0.00	0.19	4.6117
optimum kernel	7.5209	6.5772	0.00	0.34	6.8233

5 Using Error Diffusion Displacement for Optimum Kernel Design

We believe that our higher quality E_{min} error metric may be useful in kernel design, because it is the very same error that admittedly superior yet computationally expensive algorithms such as DBS try to minimise.

Our first experiment was a study of the Floyd-Steinberg-like 4-block error diffusion kernels. According to the original authors, the coefficients were found "mostly by trial and error" [4]. With our improved metric, we now have the tools to confirm or infirm Floyd and Steinberg's initial choice.

We chose to do an exhaustive study of every $\frac{1}{16}\{a, b, c, d\}$ integer combination. We deliberately chose positive integers whose sum was 16: error diffusion coefficients smaller than zero or adding up to more than 1 are known to be unstable [17], and diffusing less than 100% of the error causes important loss of detail in the shadow and highlight areas of the image.

We studied all possible coefficients on a pool of 3,000 images with an error metric E based on a standard Gaussian HVS model. E_{min} is only given here as an indication and only E was used to elect the best coefficients:



Fig. 5. Halftone of *Lena* using serpentine error diffusion (*left*) and the optimum coefficients $\frac{1}{16}\{7, 4, 5, 0\}$ (*right*) that improve on the standard Floyd-Steinberg coefficients in terms of visual quality for the HVS model used in section 3. The detail (*bottom*) shows fewer structure artifacts using optimum coefficients.

rank	coefficients	$E \times 10^4$	$E_{min} \times 10^4$
1	7 3 6 0	4.65512	3.94217
2	8 3 5 0	4.65834	4.03699
5	7 3 5 1	4.68588	3.79556
18	6 3 5 2	4.91020	3.70465
...

The exact same operation using E_{min} as the decision variable yields very different results. Similarly, E is only given here as an indication:

rank	coefficients	$E_{min} \times 10^4$	$E \times 10^4$
1	6 3 5 2	3.70465	4.91020
2	7 3 5 1	3.79556	4.68588
15	7 3 6 0	3.94217	4.65512
22	8 3 5 0	4.03699	4.65834
...

Our improved metric allowed us to confirm that the original Floyd-Steinberg coefficients were indeed amongst the best possible for raster scan. More importantly, using E as the decision variable may have elected $\frac{1}{16}\{7, 3, 6, 0\}$ or $\frac{1}{16}\{8, 3, 5, 0\}$, which are in fact poor choices.

For serpentine scan, however, our experiment suggests that $\frac{1}{16}\{7, 4, 5, 0\}$ is a better choice than the Floyd-Steinberg coefficients that have nonetheless been widely in use so far (Fig. 5).

6 Conclusion

We have disclosed an interesting property of error diffusion algorithms allowing to more precisely measure the quality of such halftoning methods. Having showed that such quality is often underestimated by usual metrics, we hope to see even more development in simple error diffusion methods.

Confirming Floyd and Steinberg's 30-year old "trial-and-error" result with our work is only the beginning: future work may cover more complex HVS models, for instance by taking into account the angular dependance of the human eye [18]. We plan to use our new metric to improve all error diffusion methods that may require fine-tuning of their propagation coefficients.

References

1. Bayer, B.: Color imaging array. U.S. patent 3,971,065 (1976)
2. Ulichney, R.A. (Digital Equipment Corporation), Void and cluster apparatus and method for generating dither templates. U.S. patent 5,535,020 (1992)

3. Ancin, H., Bhattacharjya, A., Shu, J. (Seiko Epson Corporation), Void-and-cluster dither-matrix generation for better half-tone uniformity. U.S. patent 6,088,512 (1997)
4. Floyd, R.W., Steinberg, L.: An adaptive algorithm for spatial grey scale. Proceedings of the Society of Information Display 17, 75–77 (1976)
5. Metaxas, P.: Optimal Parallel Error-Diffusion Dithering. In: Color Imaging: Device-Indep. Color, Color Hardcopy, and Graphic Arts IV, Proc. SPIE, vol. 3648, pp. 485–494 (1999)
6. Kite, T.D.: Design and Quality Assessment of Forward and Inverse Error-Diffusion Halftoning Algorithms. PhD thesis, Dept. of ECE, The University of Texas at Austin, Austin, TX (August 1998)
7. Ulichney, R.: Digital Halftoning. MIT Press, Cambridge (1987)
8. Velho, L., Gomes, J.: Digital halftoning with space-filling curves. In: Computer Graphics (Proceedings of SIGGRAPH 1991), vol. 25(4), pp. 81–90 (1991)
9. Nasanen, R.: Visibility of halftone dot textures. IEEE Trans. Syst. Man. Cyb. 14(6), 920–924 (1984)
10. Analoui, M., Allebach, J.P.: Model-based halftoning using direct binary search. In: Proc. of SPIE/IS&T Symp. on Electronic Imaging Science and Tech., San Jose, CA, February 1992, pp. 96–108 (1992)
11. McNamara, A.: Visual Perception in Realistic Image Synthesis. Computer Graphics Forum 20(4), 211–224 (2001)
12. Bhatt, et al.: Direct Binary Search with Adaptive Search and Swap, <http://www.ima.umn.edu/2004-2005/MM8.1-10.05/activities/Wu-Chai/halftone.pdf>
13. moot, <http://www.4chan.org/>
14. Wong, P.W., Allebach, J.P.: Optimum error-diffusion kernel design. In: Proc. SPIE, vol. 3018, pp. 236–242 (1997)
15. Oststromoukhov, V.: A Simple and Efficient Error-Diffusion Algorithm. In: Proceedings of SIGGRAPH 2001, in ACM Computer Graphics, Annual Conference Series, pp. 567–572 (2001)
16. Pappas, T.N., Neuhoff, D.L.: Least-squares model-based halftoning. In: Proc. SPIE, Human Vision, Visual Processing, and Digital Display III, San Jose, CA, February 1992, vol. 1666, pp. 165–176 (1992)
17. Eschbach, R., Fan, Z., Knox, K.T., Marcu, G.: Threshold Modulation and Stability in Error Diffusion. Signal Processing Magazine 20(4), 39–50 (2003)
18. Sullivan, N., Minervi, R., Pios, G.: Image halftoning using a visual model in error diffusion. J. Opt. Soc. Am. A 10, 1714–1724 (1993)

A Neuro Fuzzy Model for Image Compression in Wavelet Domain

Vipula Singh¹, Navin Rajpal², and K. Srikanta Murthy³

¹ Research Scholar GGSIPU, New Delhi

Assistant Professor ECE dept PESIT Bangalore

vipulasingh@yahoo.com

² Professor GGSIPU, New Delhi

³ Professor IS Dept PESIT Bangalore

Abstract. Image compression forms the backbone for several applications such as storage of images in a database, picture archiving, TV and facsimile transmission, and video conferencing. Compression of images involves taking advantage of the redundancy in data present within an image. This work evaluates the performance of an image compression system based on fuzzy vector quantization, wavelet based sub band decomposition and neural network. Vector quantization is often used when high compression ratios are required. The implementation consists of three steps. First, image is decomposed into a set of sub bands with different resolution corresponding to different frequency bands. Different quantization and coding schemes are used for different sub bands based on their statistical properties. At the second step, the wavelet coefficients corresponding to lowest frequency band are compressed by differential pulse code modulation (DPCM) and the coefficients corresponding to higher frequency bands are compressed using neural network. The result of the second step is used as input to fuzzy vector quantizer. Image quality is compared objectively using mean squared error and PSNR along with the visual appearance. The simulation results show clear performance improvement with respect to decoded picture quality as compared to other image compression techniques.

1 Introduction

Interest in digital image compression techniques, dates back to a few decades ago. Here we are concerned with minimizing the number of bits required to represent an image primarily for achieving information transmission and storage efficiency. All image compression algorithms strive to remove statistical redundancy and exploit perceptual irrelevancy while reducing the amount of data as much as possible. Over the last few decades researches have proposed many competing techniques such as prediction coders, transform coders, vector quantizers, trellis-coded quantizers and fractal image representation. However, due to the nature of each scheme every algorithm has its own advantages and disadvantages. Among all these schemes, however, the discrete cosine transform used in the JPEG standard has the advantage that it is well understood and it is quite mature. On the other hand vector quantization is a more recently developed technique. VQ performance is directly proportional to the codebook size and the vector size According to Shannon's rate distortion theory

larger vectors would result in better VQ performance. However, with increased vector size the required codebook size also increases, and that in-turn results in an exponential increase in encoding complexity. Consequently, for practical purposes one should limit the vector size despite the fact that better VQ performance is theoretically possible. To date, VQ techniques used in image compression may be classified into two broad categories, based on either a hard or soft decision in a membership function, namely, K-means and Fuzzy K-mean. The K-means algorithm is the most popular minimization technique using a descent algorithm. It is simple to implement but it depends strongly on the selection of the initial codebook. It is based on hard decisions and ignores the possibility that the training vector may belong to a different cluster. Artificial neural networks are receiving renewed attention in many fields where high computation rates are required. In the field of image processing applications can be found in concurrent parallel processing structures, in the recognition of characters or 2-D patterns, in some 2-D signal processing tasks and more recently in data compression for image coding systems.

Recent advances in signal processing tools such as wavelets opened up a new horizon in sub band image coding. Studies in wavelets showed that the wavelet transform exhibits the orientation and frequency selectivity of images [1-3]. Image coding using wavelets has been attempted by Antonini et al. [1], who used bi-orthogonal wavelets to obtain a set of bi-orthogonal sub bands. Woods and O'Neil [4] extended sub band decomposition to two-dimensional (2-D) signals and proposed a method for QMF design that eliminates possible aliasing error due to non ideal sub band filters. The original image was decomposed at different scales or sub bands of frequency using Mallat's pyramidal architecture [2], in which horizontal and vertical orientations are considered preferential. After the image was decomposed, the wavelet coefficients were vector quantized producing a set of multi-resolution codebooks. Averbuch et al. [5] proposed a similar image compression approach that combines pyramidal wavelet decomposition and vector quantization. In order to improve the quality of the reconstructed images, they used higher bit rate per pixel for shapes found to be more important. Codebook design is an essential process in lossy image compression techniques based on vector quantization [6-8]. Paliwal et al [9] modified the K mean algorithm for vector quantization for fast convergence without affecting the optimality of the codebook. Kadono et al [10] combined wavelet transform and vector quantization on color images.

Based on the concept of fuzzy sets, Ruspini [25] developed the first fuzzy clustering algorithm. Dunn [26] considered an alternative formulation of the clustering process and proposed the fuzzy k-means algorithm. Bezdek [27] extended Dunn's formulation and produced a family of fuzzy k-means algorithms, which includes Dunn's original algorithm as a special case. Sasazaki et al [28] suggested to use local fractal dimension to design a codebook and used fuzzy k means algorithm to group the training vectors into clusters.

Recently, artificial neural networks [11] are increasingly being examined and considered as possible solutions to problems and for application in many fields where high computation rates are required [12]. People have proposed several kinds of image compression methods [13]. The typical image compression methods are based on back propagation network [14-15]. The number of neurons in the hidden layer should be smaller than the number of units in the input and output layers. The image

is compressed in the hidden layers. However, the compression results are generally worse than traditional approaches [16-17]. Improved results probably cannot be achieved unless neural networks are incorporated directly into a state-of-the-art framework. Most successful modern coding techniques are built around filter banks [18-19]. There is evidence that the use of neural networks can improve the performance of sub band coders. Neural networks have been used to find vector quantization maps for wavelet coefficients [20], and recently Burges et al. [21] showed that preprocessing the detail coefficients using neural network DPCM reduced both the entropy of these bands and the length of the coded stream.

The main objective of this paper is to develop a Neuro-Fuzzy network based on wavelet transform and report its application to Image compression. The paper is organized as follows: Section 2 briefly reviews filter bank concepts for image compression and section 3 introduces the background necessary to define the desired input-output relations that are required to train the neural network with back propagation. Section 4 describes the fuzzy k means algorithm for vector quantization, and Section 5 presents an implementation of the image compression algorithm. Section 6 reports sample simulation results and section 7 provides concluding remarks.

2 Wavelet Transform

Wavelet transforms are multi resolution decompositions that can be used to analyze signals and images. They describe a signal by the power at each scale and position. Edges can be located very effectively in the wavelet transform domain. For image processing applications, one can use the hierarchical wavelet decomposition by Mallat [2]. The L and H filters are applied to the image in both the horizontal and vertical directions, and the filter outputs sub sampled by a factor of two, generating three orientation selective high-pass sub bands, HH, LH, HL, and a low-pass sub band LL. The process is then repeated on the LL band to generate the next level of the decomposition, and so on. As shown in Fig. 1, one octave of decomposition leads to four sub bands. Therefore, seven sub bands are obtained by iterating two times of such decomposition as shown in fig 2. The low-resolution version of the image is fed

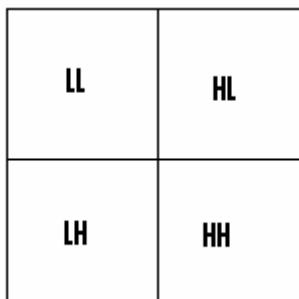


Fig. 1. First stage of a discrete wavelet transform; The image is divided into 4 sub bands using separable filters

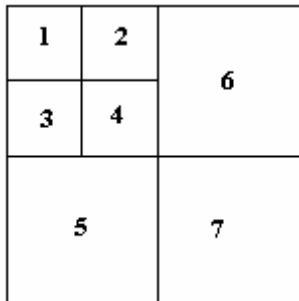


Fig. 2. Two stages of pyramidal decomposition leading to 7 octave sub bands

back to the input to produce one low-resolution version of the image and six sub bands which contain image details. The wavelet decomposition is an alternative representation of images. To compress an input image data using transform, we have to decide which coefficients to send and how many bits to encode them. Our compression algorithm consists of taking the low pass subband in full, and then deciding which coefficients within the remaining subbands to keep. Hence, the basic idea is this: the smaller the support of the wavelet, the less nonzerowavelet coefficients will correspond to an edge, so the more efficient will be our compression scheme.

3 Neural Network for Image Compression

In this section, we briefly review the basic idea of using a back propagation network to achieve image compression. A number of researchers [14-15] have shown that multilayer perceptron networks are able to learn a transform for reducing signal redundancy, and are capable of learning a reverse transform to recover the information (with some degradation) from a more compact (compressed) form of representation. In a multilayer configuration, the outputs of the units in one layer form the inputs to the next layer. The inputs to the first layer are considered as the network

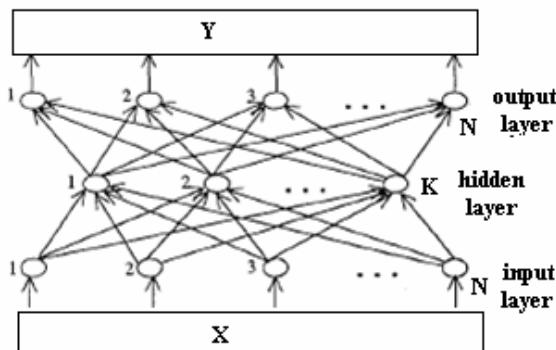


Fig. 3. A multi layered neural network

inputs, and outputs of the last layer are the network outputs. The weights of the network are usually computed by training the network using the backpropagation algorithm. The backpropagation algorithm is a supervised learning algorithm which performs a surface to arrive at a gradient descent on a squared error energy minimum. The key to the use of this method on a multilayer perceptron is the calculation of error values for the hidden units by propagating the network. The network shown in Fig. 3 has N input nodes, K hidden layer nodes ($K < N$) and N output nodes. The image compression in such a neural network proceeds in two stages; training and encoding. At the training stage, a set of input images go through the network one block by one block to allow the neural network to learn the best possible coupling weights for representation of the code-book.

After training, the network is ready for operational use. The wavelet coefficients are presented to the network one at a time. Following that, encoding begins by using the closest neuron weight to represent each individual block of the input image. The outputs of the hidden layer nodes constitute the compressed features of an input block. To achieve compression in a practical sense, the outputs of the hidden layer nodes are quantized.

4 Quantization

The process of quantization maps a signal into a series of K discrete messages. For the k^{th} message, there exists a pair of thresholds t_k and t_{k+1} and an output value q_k such that $t_k \leq q_k \leq t_{k+1}$. For a given set of quantization values, the optimal thresholds are equidistant from the values. The concept of quantizing data can be extended from scalar or one-dimensional data to vector data of arbitrary dimension. Instead of output levels, vector quantization (VQ) employs a set of representation vectors (for the one-dimensional case) or matrices (for the two dimensional case) [6-8]. The set is referred to as the “codebook” and the entries as “codewords. Hard k-means algorithm executes a sharp classification, in which each object is either assigned to a class or not. The application of fuzzy sets in a classification function causes the class membership to become a relative one and an object can belong to several classes at the same time but with different degrees [24].

4.1 K-Means Algorithm

K-means clustering, is also known as c-means clustering. The k-means algorithm partitions a collection of N vector into c groups (clusters G_i , $i=1,..,c$). The aim of that algorithm finding cluster centers (centroids) for each group. The algorithm minimizes a dissimilarity (or distance) function which is given in Equation 1.

c_i is the centroid of cluster i ; $d(x_k - c_i)$ is the distance between i^{th} centroid (c_i) and k^{th} data point; For simplicity, the Euclidian distance is used as dissimilarity measure and overall dissimilarity function is expressed as in Equation 2.

$$J = \sum_{i=1}^c J_i = \sum_{i=1}^c \sum_{k, x_k \in G_i} d(x_k - c_i) \quad (1)$$

$$J = \sum_{i=1}^c J_i = \sum_{i=1}^c \left(\sum_{k, x_k \in G_i} \|x_k - c_i\|^2 \right) \quad (2)$$

Partitioned groups can be defined by an $c \times n$ binary membership matrix (U), where the element u_{ij} is 1 if the j th data point x_j belongs to group i , and 0 otherwise. This explanation is formulated in Equation 3.

$$u_{ij} = \begin{cases} 1 & \text{if } \|x_j - c_i\|^2 \leq \|x_j - c_k\|^2, \text{ for each } k \neq i, \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Since a data point can only be in a group, the membership matrix (U) has two properties which are given equation 4 and equation 5.

$$\sum_{i=1}^c u_{ij} = 1, \forall j = 1, \dots, n \quad (4)$$

$$\sum_{i=1}^c \sum_{j=1}^n u_{ij} = n \quad (5)$$

Centroids are computed as the mean of all vectors in group i :

$$c_i = \frac{1}{|G_i|} \sum_{k, x_k \in G_i} x_k \quad (6)$$

$|G_i|$ is the size of G_i . The k-means algorithm [23] determines the following steps with a data set x_j , $j=1, \dots, n$:

Step 1. Initialize the centroids $c_i, i=1, \dots, c$. This is typically achieved by randomly selecting c points from among all of the data points.

Step 2. Determine the membership matrix U by Equation 3.

Step 3. Compute the dissimilarity function by using Equation 2. Stop if its improvement over previous iteration is below a threshold.

Step 4. Compute new centroids using by Equation 6. Go to step 2.

While the k-means algorithm converges to a local minimum, it is not guaranteed to reach the global minimum. In addition, the algorithm is very sensitive to the initial codebook. Furthermore, the algorithm is slow since it requires an exhaustive search through the entire codebook on each iteration.

4.2 Fuzzy C-Means Clustering

Fuzzy C-means Clustering(FCM), is also known as Fuzzy ISODATA, is an clustering technique which is separated from hard k-means that employs hard partitioning. The

FCM employs fuzzy partitioning such that a data point can belong to all groups with different membership grades between 0 and 1. FCM is an iterative algorithm. The aim of FCM is to find cluster centers (centroids) that minimize a dissimilarity function.

To accommodate the introduction of fuzzy partitioning, the membership matrix (U) is randomly initialized according to Equation 7. The dissimilarity function which is used in FCM is given Equation 8

$$\sum_{i=1}^c u_{ij} = 1, \forall j = 1, \dots, n \quad (7)$$

$$J(U, c_1, c_2, \dots, c_c) = \sum_{i=1}^c J_i = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m d_{ij}^2 \quad (8)$$

u_{ij} is between 0 and 1; c_i is the centroid of cluster i ; d_{ij} is the Euclidian distance between i_{th} centroid(c_i) and j_{th} data point; $m \in [1, \infty]$ is a weighting exponent. To reach a minimum of dissimilarity function there are two conditions. These are given in Equation 9 and Equation 10. Detailed algorithm of fuzzy c -means proposed by Bezdek[24]. This algorithm determines the following steps [23].

$$c_i = \frac{\sum_{j=1}^n u_{ij}^m x_j}{\sum_{j=1}^n u_{ij}^m} \quad (9)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \quad (10)$$

Step 1. Randomly initialize the membership matrix (U) that has constraints in Equation 7.

Step 2. Calculate centroids (c_i) by using Equation 9.

Step 3. Compute dissimilarity between centroids and data points using equation 8. Stop if its improvement over previous iteration is below a threshold.

Step 4. Compute a new U using Equation 7 Go to Step 2.

By iteratively updating the cluster centers and the membership grades for each data point, FCM iteratively moves the cluster centers to the "right" location within a data set. FCM does not ensure that it converges to an optimal solution. Because of cluster centers (centroids) are initialize using U that randomly initialized.(Equation 14). Performance depends on initial centroids. For a robust approach there are two ways which is described below.

1) Using an algorithm to determine all of the centroids. (for example: arithmetic means of all data points)

2) Run FCM several times each starting with different initial centroids.

5 Proposed Method

Fig 4 shows the block diagram of complete image compression system. First the image is decomposed using wavelet transform. Because the human visual system has different sensitivity to different frequency components, the following scheme is adapted. The lowest frequency band, band-1 (fig 2) is encoded with DPCM. After that these coefficients are scalar quantized. The remaining frequency bands are coded using neural network. Band-2 and 3 contain the same frequency contents for different orientation. So same neural network is used to compress the data in these bands and different neural network is used for band 5 and 6. Band-4 coefficients are coded using separate neural networks as frequency characteristics of these bands does not match with other bands. Band-7 information is discarded as it contains little information to contribute to the image from the stand this band can be assumed to be zero with little effect on the quality of reconstructed image. The output of the hidden layer of neural network is then quantized. Finally these quantized values are entropy encoded. Huffman encoding is used here. This scheme was originally proposed in [20].

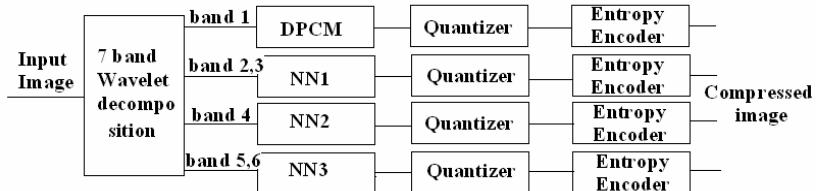


Fig. 4. Complete Image compression

6 Experimental Results and Discussion

This section presents the experimental evaluation of the proposed algorithm based on several experiments involving real image data. The set of experiments evaluate the effect of different wavelet filters on the quality of the reconstructed image. Experiments were conducted using the image ‘lena’, of size 256 x 256, with $2^8 = 256$ gray levels. Images were decomposed using Daubechies’ 1- coefficient filter (DAUB 1), 2- coefficient filter (DAUB 2), 4- coefficient filter (DAUB 4), 6- coefficient filter (DAUB 6) and 18 coefficient filter (DAUB 18), coiflet 5 coefficient filter and biorthogonal 9 coefficient filter. Band -1 is coded using DPCM, Band -2 and 3 is coded using a neural network with eight units in the input and the output and 6 hidden units i.e 8-6-8 neural network. Band -4 is coded using a 8-4-8 neural network and band -5 and 6 using 16-1-16 network. In different sets of experiments, scalar quantization (SQ) and vector quantization (VQ). was used on the coefficients of hidden layer. In the end, the coefficients are Huffman encoded.

6.1 Image Quality Evaluation

The image quality can be evaluated objectively and subjectively. A standard objective measure of image quality is reconstruction error. Two of the error metrics used to

compare the various image compression techniques are the mean square error (MSE) and the Peak Signal to Noise Ratio (PSNR). The MSE is the cumulative squared error between the compressed and the original image, whereas PSNR is a measure of the peak error. The mathematical formulae are

$$MSE = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (I_{ij} - I'_{ij})^2 \quad (11)$$

$$PSNR = 20 * \log10 (255 / \text{sqrt}(MSE)) \quad (12)$$

where $I(x,y)$ is the original image, $I'(x,y)$ is the approximated version (which is actually the decompressed image) and M,N are the dimensions of the images, 255 is the peak signal value. A lower value for MSE means lesser error, and as seen from the inverse relation between the MSE and PSNR, this translates to a high value of PSNR. Logically, a higher value of PSNR is good because it means that the ratio of Signal to Noise is higher. Subjective quality is measured by psychophysical tests or questionnaires with numerical ratings.

Table 1. MSE and PSNR of training image ‘lena’ using fuzzy k-mean algorithm with $\lambda = 5$ with different wavelet filters

Filter Type	Filter Length	MSE with FVQ	PSNR with FVQ	Compression	
				Huffman Coded	Without Huff
Daubechies	1	160.77	26.06	0.1385	0.1879
Daubechies	2	101.73	28.056	0.1393	0.1879
Daubechies	4	87.95	28.689	0.1395	0.1879
Daubechies	6	83.62	28.903	0.1382	0.188
Daubechies	18	76.15	29.314	0.1380	0.2
Coiflet	5	71.86	29.566	0.1386	0.2
Biorthogonal	9	83.99	28.97	0.1395	0.198

Table 2. PSNR of training image ‘lena’ using VQ and FVQ $\lambda = 5$ with different wavelet filters

Filter Type	Filter Length	PSNR With VQ	Compression with VQ		PSNR With FVQ	Compression with FVQ	
			Huffman Coded	Without Huff		Huffman Coded	Without Huff
Daubechies	1	26.144	0.1378	0.1879	26.06	0.1385	0.1879
Daubechies	2	28.149	0.1390	0.1879	28.056	0.1393	0.1879
Daubechies	4	28.713	0.1382	0.1879	28.689	0.1395	0.1879
Daubechies	6	29.001	0.1378	0.188	28.903	0.1382	0.188
Daubechies	18	29.402	0.1379	0.200	29.314	0.1380	0.200
Coiflet	5	29.608	0.1378	0.200	29.566	0.1386	0.200
Biorthogonal	9	29.080	0.1382	0.198	28.97	0.1395	0.198

Table 3. PSNR of training image ‘lena’ reconstructed by db18and fuzzy k-mean algorithm with different integer values of λ

Filter Type	Filter Length	PSNR With FVQ					Compression with FVQ	
		$\lambda=1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 4$	$\lambda = 5$	Huffman Coded	Without Huff
Daubechies	1	25.384	25.862	26.066	26.097	26.060	0.1378	0.1879
Daubechies	2	27.017	27.728	28.059	28.059	28.056	0.1390	0.1879
Daubechies	4	27.614	28.378	28.706	28.592	28.689	0.1382	0.1879
Daubechies	6	27.356	28.495	28.87	28.823	28.903	0.1378	0.188
Daubechies	18	27.599	28.498	29.182	29.275	29.314	0.1379	0.200
Coiflet	5	27.904	29.001	29.492	29.576	29.566	0.1378	0.200
Biorthogonal	9	27.737	28.480	28.875	28.962	28.97	0.1382	0.198



(a) Original (b) PSNR= 26.06 db (c) PSNR= 28.056 db (d) PSNR= 28.689 db



(e) PSNR= 28.903 db (f) PSNR= 29.314 db (g) PSNR= 29.566 db (h) PSNR= 28.97 db

Fig. 5. Results of ‘lena’ using fuzzy k-mean algorithm with $\lambda = 5$ with different wavelet filters
 (a) Original ‘Lena’ image (b) Reconstructed image with db1 (c) Reconstructed image with db2
 (d) Reconstructed image with db4 (e) Reconstructed image with db6 (f) Reconstructed image with db18 (g) Reconstructed image with coiflet 5 (h) Reconstructed image with biorthogonal 9



(a) $\lambda=1$ (b) $\lambda=2$ (c) $\lambda=3$ (d) $\lambda=4$ (e) $\lambda=5$

Fig. 6. ‘lena’ reconstructed by db18and fuzzy k-mean algorithm with different integer values of λ of λ

The first set of experiments evaluated the performance the proposed algorithm with distinct QMF filters for the decomposition of the image, which raises the question of which filter is the best for our compression method. Table I presents the level of compression achieved by applying different types of filters on the training image ‘lena’ and fuzzy vector quantization with $\lambda=5$. The compression achieved is described in two ways: the compression achieved after applying Huffman coding on the resulted Vector Tables and without using the Huffman coding. Huffman coding, which is a lossless compression, increases the compression by 30%. This work uses Daubechies’ wavelet of length 18 which yields the best PSNR as compared to other filters and its visual quality is better. The results are shown in Fig 9.

The second set of experiments compared the codebook design with VQ and FVQ. The wavelet coefficients in the hidden layer of the neural network can be quantized, among other methods, by VQ or FVQ. Comparing VQ versus FVQ should take into consideration two aspects : compression ratio achieved and its quality. Table II compares VQ and FVQ with the length of each vector as 16 and λ as 5 for FVQ on the training image ‘lena’. It is observed that by using FVQ, though the SNR the remains, the visual quality of the reconstructed image is better in comparison with VQ. The results are shown in Fig 10. The last set of experiments evaluated the quality of the reconstructed images with m. FVQ parameter m as evaluated as $m = 1 + 1/\lambda$, where $\lambda = 1, 2, 3, 4, 5$. Since the compression of images based on vector quantization is eventually based on crisp decisions, the quality of the codebook provided by fuzzy k mean algorithm depends strongly on the proximity of this parameter to unity. According to table IV & V, PSNR & the visual quality of the image increases as the parameter m approaches unity, with maximum SNR corresponding to $m=1.2$ in most cases. The results of this set of experiments are summarized in Fig 11.

7 Conclusion

This paper presented a neuro-wavelet based approach using fuzzy vector quantization for image compression. Compared to the neural network applied on the original image wavelet based decomposition improved the quality of reconstructed images. Among the various wavelet filters tested in the experiments, DAUB 18 resulted in slightly better results. Fuzzy vector quantization makes use of uncertainty to the benefit of the clustering process. Use of fuzzy vector quantization on the hidden layer coefficients further improves the visual quality of the reconstructed image, but it is computationally more demanding than the vector quantization algorithm.

References

- [1] Antonini, M., Barlaud, M., Mathieu, P., Daubechies, I.: Image coding using wavelet transform. *IEEE Trans. Image Processing* 1, 205–220 (1992)
- [2] Mallat, S.G.: A theory for multi resolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Machine Intell.* 11, 674–693 (1989)
- [3] Mallat, S.G.: Multi frequency channel decomposition of images and wavelet models. *IEEE Trans. Acoust., Speech, Signal Processing* 37, 2091–2110 (1989)

- [4] Woods, J.W., O'Neil, S.D.: Sub band coding of images. *EEE Trans. Acoust., Speech, Signal Processing* 34, 1278–1288 (1986)
- [5] Averbuch, A., Lazar, D., Israeli, M.: Image compression using wavelet transform and multiresolution decomposition. *IEEE Trans. Image Processing* 5, 4–15 (1996)
- [6] Gersho, A., Gray, R.M.: *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Dordrecht (1992)
- [7] Gray, R.M.: Vector quantization. *IEEE Acoust., Speech, Signal Processing Mag.* 1, 4–29 (1984)
- [8] Nasrabadi, N.M., King, R.: Image coding using vector quantization: A review. *IEEE Trans. Commun.* 36, 957–971 (1988)
- [9] Paliwal, K., Ramasubramian, V.: Comments on modified K means algorithm for vector quantizer design. *IEEE trans Image processing* 9(11), 1964–1967 (2000)
- [10] Kadono, S., Tahara, O., Okamoto, N.: Encoding of color still pictures wavelet transform and vector quantization. In: Canadian Conference on Electrical and Computer Engineering, vol. 2, pp. 931–936 (2001)
- [11] Lippmann, R.P.: An introduction to computing with neural network, pp. 36–54. IEEE Computer Society Press, Los Alamitos (1987)
- [12] Polycarpou, M.M., Ioannou, P.A.: Learning and Convergence Analysis of Neural Type Structured Networks. *IEEE Transactions on Neural Network* 3(2), 39–50 (1992)
- [13] Rao, K.R., Yip, P.: *Discrete Cosine Transform - Algorithms, Advantages, Applications*. Academic Press, London (1990)
- [14] Sonehara, N., Kawato, M., Miyake, S., Nakane, K.: Image Data Compression Using a Neural Network Model. In: IJCNN con. IEEE cat. No. 89CH2765-6, vol. 2, pp. 35–41 (1989)
- [15] Setiono, R., Lu, G.: Image Compression Using a Feedforward Neural Network. In: IJCNN, pp. 4761–4765 (1994)
- [16] Chen, W.H., Smith, C.H.: Adaptive Coding of Monochrome and Color Images. *IEEE Trans. Commun., COM-25*, 1285–1292 (1977)
- [17] Gimlett, J.I.: Use of Activity Classes in Adaptive Transform Image Coding. *IEEE Trans. Commun., COM- 23*, 785–786 (1975)
- [18] Antonini, M., Barlaud, M., Mathieu, P., Daubechies, I.: Image coding using wavelet transform. *IEEE Trans. Image Processing* 1, 205–220 (1992)
- [19] Shapiro, J.M.: Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans. Signal Processing* 41, 3445–3462 (1993)
- [20] Denk, T., Parhi, K.K., Cherkassky, V.: Combining neural networks and the wavelet transform for image compression. In: Proc. ICASSP IEEE Int. Conf. Acoust., Speech, Signal Processing, Minneapolis, Minn., pp. I-637–I-640 (1993)
- [21] Burges, J.C., Simard, Y.P., Malvar, H.S.: Improving wavelet compression with neural networks. In: Proc. Data Compression Conf., Snowbird, Utah, March 27–29, p. 486 (2001)
- [22] Max, J.: Quantizing for minimum distortion. *IEEE Trans. Inform.Theory*, 7–12 (1960)
- [23] Jang, J.-S.R., Sun, C.-T., Mizutani, E.: *Neuro-Fuzzy and Soft Computing*, pp. 426–427. Prentice-Hall, Englewood Cliffs (1997)
- [24] Bezdek, J.C.: *Pattern Recognition with Fuzzy Objective Function Algorithms*, New York, Plenum
- [25] Ruspini, E.H.: A New Approach to Clustering. *Information and Control* 15, 22–32 (1969)
- [26] Dunn, J.C.: A Fuzzy Relative of the ISODATA Process and its use in Detecting Compact Well Separated Clusters. *Journal of Cybernetics* 3(3), 32–57 (1973)

- [27] Bezdek, J.C., Ehrlich, R., Full, W.: FCM: The Fuzzy c-Means Clustering Algorithm. In: Computers and Geosciences, vol. 10(2-3), pp. 191–203 (1984)
- [28] Sasazaki, K., Ogasawara, H., Saga, S., maeda, J., Suzuki, Y.: Fuzzy vector Quantization of Images based on local fractal dimension. In: IEEE international conference on Fuzzy Systems, Canada, pp. 16–21 (2006)
- [29] Liu, H., Zhai, I., Gao, Y., Li, W., Zhou, J.: Image compression based on biorthogonal wavelet transform. In: Proceedings of ISCIT, pp. 578–581 (2005)
- [30] Premaraju, S., Mitra, S.: Efficient image coding using multi resolution wavelet transform and vector quantization. Image analysis and interpretation, pp. 135-140 (1996)

A People Counting System Based on Dense and Close Stereovision

Tarek Yahiaoui¹, Cyril Meurie¹, Louahdi Khoudour¹, and François Cabestaing²

¹ French National Institute for Transport and Safety Research (INRETS-LEOST)

20 rue Elisee Reclus, BP 317, F-59666 Villeneuve d'Ascq Cedex

{tarek.yahaoui, cyril.meurie, louahdi.khoudour}@inrets.fr

² University of Sciences and Technology of Lille, (LAGIS laboratory, UMR 8146)
Cite Scientifique, F-59655 Villeneuve d'Ascq Cedex

fcab@ieee.org

Abstract. We present in this paper a system for passengers counting in buses based on stereovision. The objective of this work is to provide a precise counting system well adapted to buses environment. The processing chain corresponding to this counting system involves several blocks dedicated to the detection, segmentation, tracking and counting. From original stereoscopic images, the system operates primarily on the information contained in disparity maps previously calculated with a novel algorithm. We show that one can obtain a counting accuracy of 99% on a large data set including specific scenarios played in laboratory and on some video sequences shot in a bus during exploitation period.

1 Introduction

Passengers counting is a very important need for transport operators. Indeed, they need reliable and precise counting information in order to plan and manage in the most appropriate way human, financial and material resources. In particular, the sharing of incomes between operators exploiting the same lines and the fraud rate evaluation are two objectives that require very precise counting information. In this case, the current approximations and statistics are not sufficient to achieve those goals. There are a lot systems for counting people in transit based on various technologies namely: infrared sensors, ultrasound, carpet contact, light rays. However, existing systems achieve too high error rates and can not correctly deal with complex situations like crowded ones. To face this problem, we have developed a counting system based on artificial vision and image processing and composed of several processing blocks exploiting stereovision. In this paper, we present the different components constituting the processing chain of the counting system. We also present the evaluation results of each block of the chain as well as the counting results on a dedicated database.

2 Proposed Counting System

Counting accurately passengers in a bus is particularly delicate. The use of a single camera system is not sufficient for reasons such as: occlusions, illumination changes and difficult configurations caused by crowd behavior in bus. Hence, we opted for an approach based on dense stereovision [1], [2]. This approach is less sensitive to illumination changes and could also provide the necessary information to detect, model and track objects or people. Furthermore, stereovision is more appropriate for cases where objects or people are very close to the sensor, which is indeed the case for people counting in buses where the integration environment constraint are very strict. With this technique of stereovision, the objective is to determine the information distance from the sensor to each point of the scene. This distance is inversely proportional to the disparity. We chose dense stereovision instead of sparse one because with this latter the primitives extraction is difficult for the reasons mentioned above. However, even if we chose to exploit dense stereovision, if it is possible we are going to extract specific points like edges. The basic idea of our proposed system is, from disparity maps calculated with a specific stereo-matching, to isolate and separate the passengers' heads in order to count them. To achieve that, the sensor is fixed vertically above the door of the bus. With this configuration, we lower the occlusions problems and whatever the crowd configuration, we suppose that the passengers' heads will rarely touch each other.

3 Processing Chain

The processing chain comprises four blocks: detection with stereo-matching, segmentation, tracking and counting. The detection block calculates, for each pair of stereoscopic images, a dense disparity map which is converted into a height map. On each map are represented distances from the ground of each point of the scene. The height maps are segmented in order to highlight the passengers' heads at different levels (adults, teenagers, children...). The results of this step are binary images containing information related to the heads; we call them "kernels". The extraction block assigns a number of parameters to the kernel: size of the kernel, shape, average greylevel, average height level... Then, with the previous information on the kernels, a tracking procedure is applied to analyze their trajectories.

4 Disparity Maps Calculation

In the literature, for the stereo-matching procedure, there are three classes of techniques of mapping: 1) The global methods: dynamic programming [3], graph theory [4], nonlinear diffusion [5] and spread belief [6]. 2) The local methods: correlation and differential approaches [7]. 3) The cooperative methods [8].

However, the counting application requires a swift technique matching. The SAD algorithm (sum of absolute differences) [9], [10], which is a technique based on the dissimilarity function between the neighborhoods of homologous pixels is an approach that allows a compromise between robustness and processing time. Let us recall that in the context of passengers counting in bus, the sensor is rather close to the observed objects. That is why, we speak about "close stereovision". The main difficulty with this geometrical configuration is the high number of occlusions in stereoscopic images which correspond to regions present in one image and not in the other. To solve this problem, we propose a technique based on a dissimilarity measure called SAD, in which we have added additional constraints.

4.1 Similarity Constraint and Dissimilarity Measurement Weighting

We call similarity constraint any discriminating similarity criterion which may exist between the pixel to match and its homologous or between their respective neighborhoods. The objectives of the use of this kind of constraints are: 1) Reduction of the computation time by rejecting candidates pixels that do not verify the constraints. 2) Increase the accuracy of the matching by choosing pixels strongly correlated.

In our case, we want to improve the stereo-matching quality. Therefore, we propose to refine the selection of homologous pixels by weighing the SAD dissimilarity measure. This is done by introducing a weighting factor whose value depends on the verification of the similarity criterion. When the similarity criterion is verified, the weighting procedures consists in reducing the dissimilarity measure in order to match only the pixels verifying the similarity criterion. The value that the weight can take when the similarity criterion is not verified, does not affect in any way the dissimilarity measure value. For modeling the influence of a constraint, we chose to multiply this similarity criterion by a weighting factor. The dissimilarity measurement is written:

$$C_{SAD}(x, y, s) = \sum |G(x + i + s, y + j) - D(x + i, y + j)| \quad (1)$$

$G(x,y)$: The greylevel of the pixel (x,y) to match belonging to the left image.
 $D(x,y)$: The greylevel of the pixel (x,y) in the right image. S : The gap between the two pixels (left and right).

The dissimilarity measure after the introduction of the similarity criterion is written as follows:

$$C_{sim}(x, y, s) = coef \sum |G(x + i + s, y + j) - D(x + i, y + j)| \quad (2)$$

with $coef$, the weighting factor. $Coef = 1$ if similarity criterion is not verified. $Coef = coef_0$ and $0 < coef_0 < 1$ otherwise.

4.2 Proposed Constraints

We have identified four similarity constraints detailed below. Thus, improvement brought by the introduction of the constraints, in terms of accurate matching, does not require a huge increase in the processing time. 1) Similarity of the neighborhoods' centers greylevels. 2) Similarity of the type of pixels (belonging to edges or not). 3) Similarity of greylevels profiles corresponding to the Centerlines of Calculation Neighborhoods (called CCN in the next equation). 4) Similarity of the type of regions including the pixels to match (region with motion or static one). We have respectively associated the coefficients α , β , γ and μ with these constraints. The values of these coefficients will vary depending on whether the constraints are verified or not.

$$\alpha = \begin{cases} 1 & \text{if } G \text{ and } D \text{ do not have the same greylevels} \\ \alpha_0 & \text{with } 0 < \alpha_0 < 1 \text{ otherwise} \end{cases} \quad (3)$$

$$\beta = \begin{cases} 1 & \text{if } G \text{ and } D \text{ do not represent edges} \\ \beta_0 & \text{with } 0 < \beta_0 < 1 \text{ otherwise} \end{cases} \quad (4)$$

$$\gamma = \begin{cases} 1 & \text{if the CCN do not have the same greylevels profile} \\ \gamma_0 & \text{with } 0 < \gamma_0 < 1 \text{ otherwise} \end{cases} \quad (5)$$

$$\mu = \begin{cases} 1 & \text{if } G \text{ and } D \text{ do not correspond to a mobile region} \\ \mu_0 & \text{with } 0 < \mu_0 < 1 \text{ otherwise} \end{cases} \quad (6)$$

With G corresponding to $G(x+s,y)$, and D to $D(x,y)$. We have determined the optimal values of α_0 , β_0 , γ_0 and μ_0 experimentally by choosing the values that minimize the stereo-matching error rate according to a given ground truth proposed by SCHARSTEIN AND SZELISKI.

4.3 Constraints Association

So far, we have submitted four similarity constraints that we use to improve the accuracy of matching. Knowing that each of these constraints is of a different nature, it becomes interesting to combine them and analyze their complementarities. The main idea is to gather as much information as possible on the pixel to match and its homologous and also on their neighborhoods. We chose to use an additive model for the calculation of the dissimilarity , which is to add the dissimilarity of the four criteria. The global formulation becomes:

$$C(x, y, s) = (\alpha + \beta + \gamma + \mu) \sum |G(x + i + s, y + j) - D(x + i, y + j)| \quad (7)$$

To evaluate our approach we compared it to SAD (Some of absolute differences), SSD (Some of squared differences), ZSAD (Zero mean some of absolute differences) and ZSSD (Zero mean some of squared differences) on static data with

ground truths proposed by SCHARSTEIN AND SZELISKI and dynamic data with ground truth by using the VANDERMARK sequence [11]. By comparing the stereo-matching error rate curves of our approach and those of the others we find that our approach has an error rate which is 3% lower than those of the others [12]. This result is mainly obtained in occluded areas of the images. Figure 1 provides two disparity maps calculated on a pair of stereoscopic images. We can notice that for SAD algorithm, some matching errors appear (marked with circles). This shows visually the improvement brought by the introduction of constraints in SAD computation.

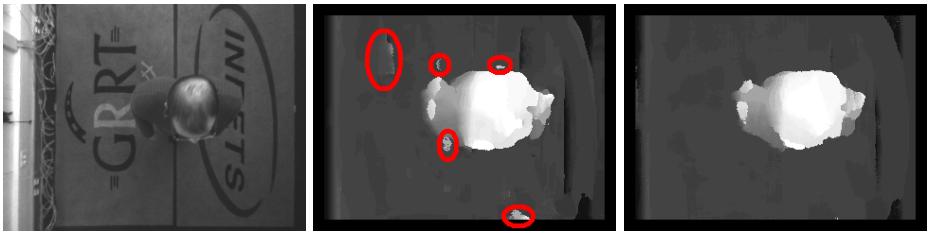


Fig. 1. Disparity maps corresponding to laboratory stereoscopic images (from left to right: initial left image, SAD disparity map with errors outlined with circles, our approach of disparity map)

5 Exploiting Disparity Maps for Tracking and Counting People

The disparity maps calculated by our stereo-matching method are transformed into height maps by simple triangulation. To highlight the passengers' heads, we propose a segmentation technique based on a given number of height intervals. For each interval, we have a binary image on which appear only regions with a height belonging to the interval. In a given interval, the application of morphological operations as openings with circular structuring elements allow us to locate in space and time, the heads of the passengers. At the end of the processing within all the intervals, we obtain a binary image containing kernels supposed to represent the heads. Figure 2 shows an example of segmentation of a height map corresponding to two persons passing under the sensor.

Then, in tracking application, we associate to each kernel a vector of attributes. This vector is a set of properties that distinguish each object from others in the same scene. In our case, the objects we need to track are the passengers' heads and attributes are actually properties that vary from one head to another. The considered attributes are: kernel size, kernel width, kernel length, average height in centimeters, average greylevel (from initial image), kernel coordinates x,y (in the binary image). We use a technique based on Kalman filtering



Fig. 2. Example of a height map segmentation (from left to right: initial left image, the height map, the kernels map)

for the kernels tracking. It yields a prediction of the kernels positions and compares them to the current attributes vectors. When implementing this technique, we have decided to simplify the calculation by limiting the prediction on the two components corresponding to the x and y coordinates of the kernel barycenter in the image. The counting technique we have developed is based on the analysis of the notion of valid trajectories [12].

6 Evaluation and Results

First, we note that the counting system has been fully assessed on real dataset. Data on which the system has been tested come from two different bases: laboratory data according to 30 specific scenario scripts provided by the Parisian Transport Operator (RATP) and 3 real video sequences acquired in an operating bus. Most of the scenarios represent exiting persons. The counting results presented in Figure 3 indicates the number of persons entering or exiting for each sequence of the laboratory. In this figure, we can see the ground truth counting results versus the counting results computed by our algorithm. One can notice that whatever the difficulty of the scenario, the difference between real counting and calculated one with our approach is very low. Indeed these differences are included in the interval $[-1; +1]$. It is also the case for the counting results of the three bus scenarios presented in Figure 4. There are no counting error in the two first scenarios and a slight under-estimation for the third scenario. This is a very encouraging result demonstrating the robustness of our algorithm which is able to cope with various situations (high density groups of persons moving in opposite directions, persons of different sizes, carrying bags...). In order to determine globally the accuracy of our counting system, that is to say considering all the scenarios and mixing entries and exits, we have defined an error rate which is calculated as follows in the formula 8. In this formula, we consider the real counting (ground truth) as basis of comparision and we determine the difference between the counting with our algorithm. Thus, the error rate is around 1%. Among the laboratory scenarios, the same error rate is obtained whatever the

illumination type considered (scenarios 1-15 with daylight and scenarios 16-30 with artificial light). When analyzing more finely the counting results, we see that our system under-estimate systematically the number of persons. Several reasons could explain this fact: 1) the difficulty to detect persons whose size is small and who can be confused with objects. 2) The size of the structuring element in the segmentation step of the disparity map. 3) the merging of two trajectories, corresponding to two different persons.

$$Error_{counting} = 100 \frac{(Real_{counting} - Automatic_{counting})}{Real_{counting}} \quad (8)$$

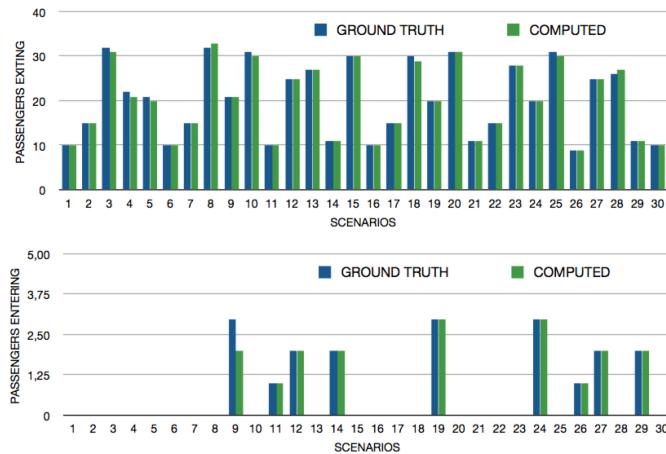


Fig. 3. Counting results for 30 scenarios in laboratory (from top to bottom: exiting from and entering in the bus by the same door)

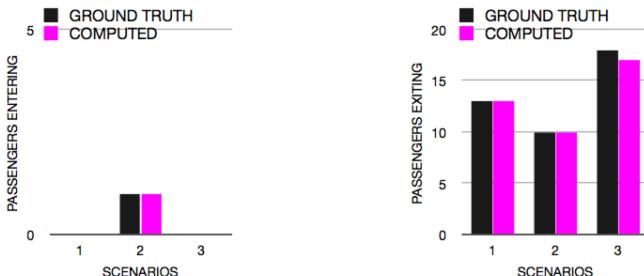


Fig. 4. Counting results for 3 scenarios in bus (from left to right: entering in and exiting from the bus by the same door)

7 Conclusion

In this work, we presented a problem that requires combination of tools emerging from different image processing disciplines. Stereovision was the most important tool we used and adapted to our application. We proposed a stereo-matching approach that integrates similarity criteria to improve the calculation of the disparity maps. It is an issue where the fundamental aspect is attached to the applicative aspect, which allows us to accomplish two objectives, namely, to produce a precise counting system with 99% of good counting and to contribute to the development of new basic tools in the stereoscopic images processing by proposing a matching approach exploiting similarity constraints. The counting accuracy must be refined in a more intensive evaluation. This is currently carried out with a new data set comprising 150 stereo sequences coming from a bus in exploitation. We have good hope to stay under 2% of error thanks to our complete processing chain.

References

1. Beyme, D.: Person counting using stereo. In: Workshop on Human Motion, pp. 127–133 (2000)
2. Terada, K., et al.: A counting method of the number of passing people using a stereo camera. In: 25th IEEE Conference of Industrial Electronics Society, pp. 338–342 (1999)
3. Bobick, A.F., Intille, H.S.: Large occlusion stereo. International Journal On Computer Vision 33, 181–200 (1999)
4. Paris, S., Sillon, F.: Optimisation à base de flot de graphe pour l'acquisition d'informations 3d à partir de séquences d'image. In: 15 èmes Journées de l'Association Française d'Informatique Graphique, pp. 165–182 (2002)
5. Mansouri, A.R., Mitiche, A.: Selective image diffusion: Application to disparity estimation. In: IEEE International Conference on Image Processing, vol. 3, pp. 284–288 (1998)
6. Sun, J., Zheng, N.N.: Stereo matching using belief propagation. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(7), 787–800 (2003)
7. Wei, G.Q., Brauer, W., Herzinger, G.: Intensity and gradient based stereo matching using hierarchical gaussian basis functions. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1143–1160 (1998)
8. Zhang, Y., Kambhamattein, C.: Stereo matching with segmentation-based cooperation. In: 7th European Conference on Computer Vision, vol. 2, pp. 556–571 (2002)
9. Martin, J., Krowley, J.L.: Experimental comparison of correlation techniques. In: International Conference on Intelligent Autonomous Systems (2002)
10. Haris, S., Vandermark, W., Cavrilà, D.M.: A comparative study of fast dense stereo vision algorithms. In: IEEE Intelligent Vehicles Symposium, pp. 319–324 (2004)
11. Vandermark, W., Gavrila, D.M.: Real-time dense stereo for intelligent vehicles. IEEE Transactions on Intelligent Transportation Systems 7(1), 38–50 (2006)
12. Yahiaoui, T.: Une approche de stéréovision dense intégrant des contraintes de similité. Application au comptage de passagers entrant et sortant d'un autobus. PhD thesis, University of Lille (2007)

Topology-Preserving Discrete Deformable Model: Application to Multi-segmentation of Brain MRI

Sanae Miri^{1,2}, Nicolas Passat¹, and Jean-Paul Armspach²

¹ LSIIT, UMR 7005 CNRS/ULP, Strasbourg 1 University, France

² LINC, UMR 7191 CNRS/ULP, Strasbourg 1 University, France

Abstract. Among the numerous 3D medical image segmentation methods proposed in the literature, very few have intended to provide topologically satisfying results, *a fortiori* for multiple object segmentation. In this paper, we present a method devoted to parallel segmentation of the main classes of cerebral tissues from 3D magnetic resonance imaging data. This method is based on a multi-class discrete deformable model strategy, starting from a topologically correct model, and guiding its evolution in a topology-preserving fashion. Validations on a commonly used cerebral image database provide promising results and justify the further development of a general methodological framework based on the concepts exposed in this preliminary work.

Keywords: Topology preservation, multi-segmentation, discrete deformable model, medical imaging.

1 Introduction

When performed on medical data, segmentation consists in decomposing the image support into two sets: the searched structure, and the “useless” part of the image, usually by using photometric and/or shape knowledge related to the anatomical elements and the image acquisition process. However complex organs, such as the brain, require more sophisticated segmentation strategies. Indeed, the brain is composed of several different tissues and anatomical structures presenting non-trivial shapes and organised in a complex fashion. Consequently, in order to provide results enabling to facilitate the analysis of brain images, segmentation methods have to correctly divide such images into $k > 2$ classes corresponding to the several anatomical and pathological elements visualised. The determination of multiple classes, and the complex shape and organisation of brain structures require to base the developed strategies not only on “classical” intensity and shape hypotheses, but also on more sophisticated geometrical, relational and topological ones. In this context, the notion of deformable model can deal with these different kinds of *a priori* anatomical knowledge, and then lead to the creation of efficient tools.

In the sequel, a method devoted to brain structure segmentation from 3D magnetic resonance imaging (MRI) data is described. It relies on a multi-class

discrete deformable model strategy, starting from a topologically-correct model, and evolving without topology modifications until segmenting the brain into four classes corresponding to the four main intracranial structures. This method constitutes a preliminary attempt to illustrate and validate the potential efficiency of a general methodology consisting in developing complex - and realistic - anatomical models of the brain and deforming them under high-level *a priori* anatomical knowledge to obtain accurate segmentation results.

2 Related Work

In [8][10], some strategies are described for vascular networks segmentation. Based on the assumption that these networks are organised as tree structures, they propose to monotonically modify a simply connected model by addition (or subtraction) of simple points [2]. Using the same idea, brain surface segmentation methods [6][5] use the assumption that the cortex is a topological hollow sphere. They thicken/thin a subset/superset corresponding to the white matter/intracranial volume to finally detect the cortex/cortical surface. Topology preserving is handled by considering simple point [6] or complexes [5].

The first methods devoted to *sequential* multi-segmentation of 3D brain images are described in [9][7]. They segment cerebral anatomical structures in a coarse to fine way, the result of each segmentation step enabling to initialise the next one. Recently, the first *parallel* cerebral structure multi-segmentation method [1] has been proposed. Based on a topological model of the brain, it iteratively performs a classification/homotopic skeletonisation/homotopic thickening process. The initialisation requires a rigid-registration of a precomputed topological model which has to present structures sufficiently large not to be topologically altered by the deformation field. Moreover, the issue of modelling and preserving the topology of a non-binary image is not clearly considered (all classes are handled in a 6-adjacency framework, and some voxels may be unclassified). Finally, the iteration of numerous steps is computationally quite expensive.

As in [1], the work described hereafter is based on a topological model of the brain structures which evolves, thanks to the notion of simple point, in a topology preserving fashion. However, it presents several major differences. The topological model is more simple than the one of [1] (both still being non-anatomically compliant) but it enables to correctly model and preserves topological properties during the deformation process. The deformation algorithm is not based on successive monotonic processes, but on a smooth non-monotonic one, leading to a low computation time method. Finally, this work constitutes a preliminary study devoted to assess the potential efficiency of the proposed methodology.

3 Background

3.1 Anatomical and Physical Notions

Brain Anatomy. The brain is mainly composed of three kinds of tissues: grey matter, white matter and cerebrospinal fluid. It is constituted of several structures

located in the intracranial volume: brainstem, cerebellum, ventricles, cortex, etc. Most of them have complex shapes, but are quite invariant from one brain to another, particularly from topological and relational points of view. In a macroscopic and simplified way, the brain is surrounded by a layer of tissues and cerebrospinal fluid. Except on its inferior face (where are located the cerebellum and the brainstem), the external surface of the brain is composed of the cortex, which is a thick convoluted ribbon of grey-matter. Under this cortical surface, white matter surrounds the ventricles, which are cavities containing cerebrospinal fluid, linked together by thin tunnels.

Magnetic Resonance Imaging. Magnetic resonance imaging (MRI) is a class of medical image acquisition protocols taking advantage of the magnetic properties of hydrogen atoms located in living tissues to visualise internal anatomical structures. There exist several modalities of MRI, each one presenting specific properties. The most commonly used for brain visualisation is T1 MRI, enabling to discriminate the white matter, the grey matter and the cerebrospinal fluid, which present decreasing and globally homogeneous intensities in such data. An example of T1 MRI axial slice is illustrated in the 1st picture of Fig. 1, where we can observe three classes of intensity corresponding to the three classes of tissues previously enumerated.

3.2 Notations and Hypotheses

Notations In the sequel, the grey matter, white matter, and cerebrospinal fluid will be denoted by GM, WM, and CSF, respectively. A 3D MRI data will be considered as a function $I : E \rightarrow \mathbb{N}$ (a MRI data having discrete positive values), where E is the support of I , and has the form $[0, d_x - 1] \times [0, d_y - 1] \times [0, d_z - 1]$, d_x, d_y, d_z , being the dimensions of I (generally 256, for brain millimetric data).

Hypotheses. The purpose of this work is the evaluation of the potential efficiency of parallel segmentation of the different cerebral structures, based on a discrete topology-guided deformable model process. In order to lead to a final application which will be fully efficient, such a strategy will require to deal with several issues, some of them still being open problems (this will be discussed in Section 6). To be able to evaluate the feasibility and efficiency of this strategy before dealing with these issues, we chose to make two main simplifying assumptions: (1) the three different tissues constituting the brain present a quite homogeneous signal, *i.e.* MRI noise and signal distortion are low, and (2) the brain presents a relational and topological structure more simple than the real one (by omission of the smallest structures), *i.e.* the brain will be considered as composed of four “tissue layers” hierarchically surrounded by each others. It has to be noticed that these simplifying hypotheses *do not alter* the general segmentation methodological framework which will be illustrated hereafter. The first hypothesis will enable to provide a simple constraint to deform the topological model during the segmentation process, while the second one will allow the development of a tractable multi-class topological model.

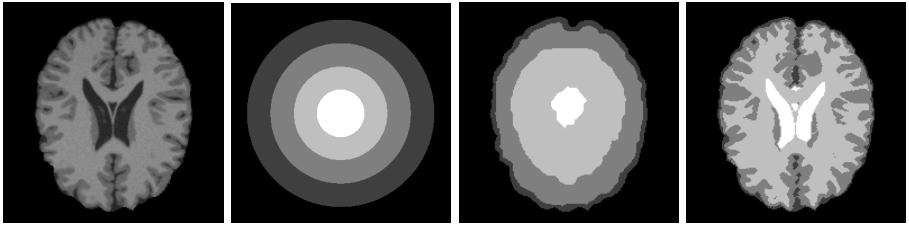


Fig. 1. From left to right: axial slice of a T1 MRI restricted to the intracranial volume ($I|_{E'}$); topological model (from white to dark grey: C_v^i , C_w^i , C_g^i , and C_s^i); initialisation of the segmentation process of the MRI (C^i); expected result of the segmentation of the MRI (from white to dark grey: C_v^i , C_w^i , C_g^i , and C_s^i)

4 Method

4.1 Input/Output

The method takes as input a T1 MRI of the brain, $I : E \rightarrow \mathbb{N}$, from which the intracranial volume $E' \subset E$ has been extracted¹ (Fig. 1, 1st picture), and two threshold values $\mu_1 < \mu_2 \in \mathbb{N}$ delimiting the T1 signal intensity between CSF/GM, and GM/WM, respectively. The method output is a partition $C = \{C_s, C_g, C_w, C_v\}$ of E' , where C_s , C_g , C_w , and C_v correspond to the sulcal (*i.e.* external) CSF, GM, WM, and ventricular (*i.e.* internal) CSF classes, respectively (Fig. 1, 4th picture).

4.2 Initialisation

The segmentation process has to be initialised from a topological model C^i of E' having the desired topology: C_v^i is simply connected (1 connected component, 0 hole, 0 cavity), and successively surrounded by C_w^i , C_g^i , and C_s^i which are topological hollow spheres (1 connected component, 0 hole, 1 cavity), hierarchically organised, as illustrated in a 2D fashion in Fig. 1 (2nd picture).

In order to generate this initial model, a distance map is computed from E' . This discrete map (providing the chamfer distance of each point of E' to its border, *i.e.* to the brain hull) is then used to guide a homotopic reduction process. Simple points are iteratively removed from E' , with a highest priority to the points being the closest from the border of E' . These removals are performed until reaching a distance $d_v > 0$ corresponding to an approximation of the sulcal CSF thickness (which can be expressed as a ratio of the highest distance of the map). The set composed by the removed points (necessarily presenting a hollow sphere topology) corresponds to the sulcal CSF class C_v^i , while the remaining set of E' corresponds to C_w^i , C_g^i , and C_s^i . The same process is iterated on this

¹ Justified by the existence of efficient tools, the preliminary extraction of the intracranial volume is a frequent preprocessing procedure before cerebral structure segmentation [5, 6, 4].

remaining set, with distances $d_g, d_w > 0$, to successively generate C_g^i (GM) and C_w^i (WM) which, by construction, also present a hollow sphere topology. The finally remaining component, thus presenting a simply-connected topology, corresponds to the “central” part of the model, *i.e.* C_s^i (ventricular CSF). A 2D axial slice of an example of initial model is illustrated in Fig. II (2nd picture).

In \mathbb{Z}^3 , the use of such a model, organised as a set of “hierarchically included” connected components, implies to choose dual adjacencies for these successive components (*i.e.* classes). The 6-adjacency has been considered for C_g^i (and thus C_v^i), since the GM is geometrically organised as a “thick” ribbon, while the 26-adjacency has been considered for C_w^i, C_v^i , since they both present “thin” details near the cortex.

From a topological point of view, C^i , although composed of four distinct classes, can be considered as a binary image made of an object $X = C_s^i \cup C_w^i$ and of the background $\bar{X} = C_g^i \cup C_v^i$, in a (26, 6)-adjacency framework. This is justified by the fact that each class is adjacent to - at most - two other classes (the surrounded and surrounding ones), which are not adjacent with each other. Both adjacent classes then locally correspond to the “background” of this class, considered as a binary structure.

4.3 Discrete Deformable Model

The discrete deformable model, which constitutes the main part of the method, consists in “deforming” the four classes without altering their topology until reaching the segmentation of the searched structures. This deformation is performed by modifying the frontiers between the classes, which is actually equivalent to modify the frontier between the sets X and \bar{X} . The - topology preserving - evolution of the frontier of a binary object can be performed by adding/removing simple points to/from this object. Adding/removing a point to/from X is equivalent to add/remove a point to/from C_s^i or C_w^i while removing/adding it from/to C_g^i or C_v^i . For a given point, the binary growth/reduction of X then corresponds to a reclassification (class modification between two “adjacent” classes) in C^i . It has to be noticed that, given the chosen dual adjacencies for the successive classes, a simple point of X or \bar{X} is only adjacent to one connected component of X and one connected component of \bar{X} , *i.e.*, only adjacent to its own class and the class where it could be reclassified. Consequently, (1) such points are located at the frontier between two classes, and (2) there is no ambiguity regarding their potential reclassification.

The proposed deformation model aims at iteratively reclassifying the points located at the frontiers between classes, making all these frontiers parallelly evolving in a smooth and non-monotonic way. The deformation is guided by photometric constraints. A cost is provided for each point $x \in E'$: if the grey-level value $I(x)$ of x is not coherent w.r.t. the expected value interval (provided by the input threshold values μ_1 and μ_2) of the class it belongs to, the - positive - distance between $I(x)$ and this interval is assigned as cost for x . The deformation model iteratively switches “misclassified” simple points from one class to another, giving the highest priority to the “most misclassified” ones (*i.e.* those having the

Algorithm 1 - Deformable model algorithm

```

repeat
  1 - Frontier point determination
   $FP_{\{s,g\}} = (C_s^i \cap N_6^*(C_g^i)) \cup (C_g^i \cap N_{26}^*(C_s^i))$ 
   $FP_{\{g,w\}} = (C_g^i \cap N_{26}^*(C_w^i)) \cup (C_w^i \cap N_6^*(C_g^i))$ 
   $FP_{\{w,v\}} = (C_w^i \cap N_6^*(C_v^i)) \cup (C_v^i \cap N_{26}^*(C_w^i))$ 
  /*  $N_k^*(A)$  is the set of points of  $\bar{A}$   $k$ -adjacent to  $A$  */
  2 - Simple point determination
   $SP_{26} = \{x \in X \mid x \text{ is 26-simple for } X\}$ 
   $SP_6 = \{x \in \bar{X} \mid x \text{ is 6-simple for } \bar{X}\}$ 
  3 - Candidate point determination
   $CP = (SP_6 \cup SP_{26}) \cap (FP_{\{s,g\}} \cup FP_{\{g,w\}} \cup FP_{\{w,v\}})$ 
  /*  $CP$ : Simple points at the frontier between two classes */
  4 - Cost evaluation
  for all  $x \in CP \cap FP_{\{s,g\}}$  (resp.  $CP \cap FP_{\{g,w\}}$ , resp.  $CP \cap FP_{\{w,v\}}$ ) do
     $v(x) = I(x) - \mu_1$  (resp.  $I(x) - \mu_2$ , resp.  $I(x) - \mu_1$ )
    if  $x \in C_g^i$  (resp.  $C_w^i$ , resp.  $C_v^i$ ) then
       $v(x) = -v(x)$ 
    end if
  end for
  /* Correctly classified points have a negative cost  $v$  */
  5 - Point selection and reclassification
  if  $\max(v(CP)) > 0$  /* with  $\max(v(\emptyset)) = -\infty$  */ then
    Let  $y \in CP$  such that  $v(y) = \max(v(CP))$ 
    Let  $C_\alpha^i \in \{C_s^i, C_g^i, C_w^i, C_v^i\}$  such that  $y \in C_\alpha^i$ 
    Let  $C_\beta^i \in \{C_s^i, C_g^i, C_w^i, C_v^i\}$  such that  $y \in FP_{\{\alpha,\beta\}}$ 
     $C_\alpha^i = C_\alpha^i \setminus \{y\}$ 
     $C_\beta^i = C_\beta^i \cup \{y\}$ 
  end if
  until  $\max(v(CP)) \leq 0$ 
   $C = C^i$ 

```

highest cost), until no simple point or no misclassified point is detected. This process, informally explained here, is clearly detailed in Alg. II

5 Experiments and Results

5.1 Complexity and Computation Time

For simplicity's sake, the segmentation algorithm detailed in Alg. II is presented in a “non-optimal” fashion, using set-based notations. However, it was implemented using efficient data structures and strategies (based on ordered FIFO lists), enabling to reach its optimal - linear - algorithmic complexity $O(|E'|)$.

The proposed method has been implemented in C++, and integrated into the multidimensional medical image processing and analysis software platform developed by the LINC. The tests and validations were performed on a 3 GHz processor / 2 GB memory personal computer. In such conditions the computation time for processing a 256^3 3D T1 MRI vary between 1 and 2 minutes,

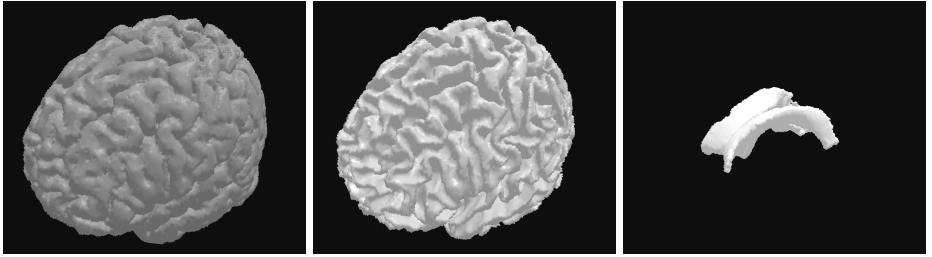


Fig. 2. 3D visualisation of brain segmentation results provided by the proposed method. From left to right: C_g (GM), C_w (WM), and C_v (ventricular CSF).

Table 1. Sensitivity, specificity and similarity measures of segmentation results on the BrainWeb database, for various noise ratios

Noise	Sensitivity			Specificity			Similarity		
	CSF	WM	GM	CSF	WM	GM	CSF	WM	GM
0%	88.5%	99.4%	81.5%	99.4%	98.2%	99.8%	83.6%	91.8%	89.0%
1%	88.9%	99.3%	83.2%	99.3%	98.5%	99.8%	83.2%	93.1%	90.0%
3%	89.7%	95.4%	85.9%	98.9%	99.6%	99.4%	79.8%	95.7%	90.4%
5%	89.8%	87.2%	81.2%	98.6%	99.7%	98.5%	76.4%	91.6%	84.8%
7%	90.0%	80.0%	75.9%	98.3%	99.6%	97.7%	74.0%	87.1%	79.2%
9%	82.3%	62.1%	66.9%	97.7%	99.6%	95.7%	65.6%	75.0%	68.1%

depending on the quality of the data. This relatively low computation time, associated to the automation of the process, results in an easy use of the method for clinicians.

5.2 Validations

In order to quantitatively and qualitatively validate the method, T1 MRI data provided by the commonly used BrainWeb² database have been considered.

Results obtained for images corresponding to noise ratios from 0% to 9% of corrupted voxels have been analysed by considering, for each class of tissue³, three statistical measures: sensitivity ($tp/(tp + fn)$), specificity ($tn/(tn + fp)$) and similarity ($2.tp/(2.tp + fp + fn)$), with tn , tp , fn and fp being the number of true negative, true positive, false negative and false positive voxels, respectively. These measures are summarised in Table 1. The result illustrated in Fig. 2 corresponds to the first line of this table.

From a quantitative point of view, the obtained results are still not perfect. For non-noisy data, they are similar to results proposed by other kinds of methods, such as the statistical one developed in [4]. For noisy data, the results are

² <http://www.bic.mni.mcgill.ca/brainweb>

³ The classes C_s and C_v , both corresponding to CSF, have been fused, since the BrainWeb ground truth does not discriminate them.

less satisfactory, since the proposed method - in its primary version proposed here - strongly relies on photometric constraints. From a qualitative point of view, the results are, however, much better than those obtained from [4], and more generally from methods which do not rely on anatomical (*i.e.* geometrical, topological and/or relational) *a priori* knowledge. As an example, the cortex presents here a real thick surface, while the ventricular CSF is fully surrounded by the grey matter.

6 Conclusion

This method and encouraging results exposed in this paper, constitute preliminary works related to the parallel topology-preserving segmentation of structures from medical data, in a discrete deformation model framework. This field of research has not been extensively studied until now, probably because it requires to find correct solutions to non-trivial theoretical and practical issues. Here, a simplifying hypothesis enabled to consider a four label image of the brain as a binary one. The real topology of brain structures will not allow to keep considering such an hypothesis. The most important issue is then the determination of satisfying solutions for defining and modelling the topology of digital images composed of more than two classes of labels, and to deform such images while leaving unchanged their topological properties. Another way for further research deals with the determination of more sophisticated constraints for guiding the deformation process. Here, photometric parameters were considered. Other kinds of *a priori* anatomical knowledge [3] should be involved in the guidance of the deformation process, in order to make it as reliable as a human expert. The main challenges will consist in determining and formalising them in a way finally enabling to obtain a method being robust, while not excessively altering its algorithmic complexity and computation time.

References

1. Bazin, P.-L., Pham, D.L.: Topology-preserving tissue classification of magnetic resonance brain images. *IEEE Transactions on Medical Imaging* 26(4), 487–496 (2007)
2. Bertrand, G., Malandain, G.: A new characterization of three-dimensional simple points. *Pattern Recognition Letters* 15(2), 169–175 (1994)
3. Bosc, M., Heitz, F., Armspach, J.-P.: Statistical atlas-based sub-voxel segmentation of 3D brain MRI. In: *Proceedings of International Conference on Image Processing - ICIP 2003. 10th International Conference, Barcelona, Spain, September 14-17 2003*, vol. 2, pp. 1077–1080. IEEE Signal Processing Society, Los Alamitos (2003)
4. Bricq, S., Collet, C., Armspach, J.-P.: Triplet Markov chain for 3D MRI brain segmentation using a probabilistic atlas. In: *Proceedings of International Symposium on Biomedical Imaging - ISBI 2006. 3rd International Symposium, Arlington, VA, USA, April 6-9 2006*, pp. 386–389 (2006)

5. Cointepas, Y., Bloch, I., Garnero, L.: Joined segmentation of cortical surface and brain volume in MRI using a homotopic deformable cellular model. In: Proceedings of 3D Digital Imaging and Modeling - 3DIM 1999, 2nd International Conference, Ottawa, Canada, October 4-8 1999, pp. 240–248 (1999)
6. Daragon, X., Couprie, M.: Neo-cortex segmentation from IRM data in order topology. In: Proceedings of Reconnaissance de Formes et Intelligence Artificielle - RFIA 2002, 13th Francophone Conference, Angers, France, January 8-10, 2002, vol. 3, pp. 809–818 (2002)
7. Dokládal, P., Bloch, I., Couprie, M., Ruijters, D., Urtasun, R., Garnero, L.: Topologically controlled segmentation of 3D magnetic resonance images of the head by using morphological operators. *Pattern Recognition* 36(10), 2463–2478 (2003)
8. Dokládal, P., Lohou, C., Perrotin, L., Bertrand, G.: Liver blood vessels extraction by a 3-D topological approach. In: Taylor, C., Colchester, A. (eds.) MICCAI 1999. LNCS, vol. 1679, pp. 98–105. Springer, Heidelberg (1999)
9. Mangin, J.-F., Frouin, V., Bloch, I., Régis, J., López-Krahe, J.: From 3D magnetic resonance images to structural representations of the cortex topography using topology preserving deformations. *Journal of Mathematical Imaging and Vision* 5(4), 297–318 (1995)
10. Passat, N., Ronse, C., Baruthio, J., Armspach, J.-P., Bosc, M., Foucher, J.: Using multimodal MR data for segmentation and topology recovery of the cerebral superficial venous tree. In: Bebis, G., Boyle, R., Koracin, D., Parvin, B. (eds.) ISVC 2005. LNCS, vol. 3804, pp. 60–67. Springer, Heidelberg (2005)

A Real Time Human Detection System Based on Far Infrared Vision^{*}

Yannick Benezeth¹, Bruno Emile¹, Hélène Laurent¹,
and Christophe Rosenberger²

¹ Institut Prismé, ENSI de Bourges - Université d'Orléans - 88 boulevard Lahitolle,
18020 Bourges cedex - France

² Laboratoire GREYC, ENSICAEN - Université de Caen - CNRS, 6 boulevard
Maréchal Juin, 14000 Caen - France
yannick.benezeth@ensi-bourges.fr

Abstract. We present in this article a human detection and tracking algorithm using infrared vision in order to have reliable information on a room occupation. We intend to use this information to limit energetic consumption (light, heating). We perform first, a foreground segmentation with a Gaussian background model. A tracking step based on connected components intersections allows to collect information on 2D displacements of moving objects in the image plane. A classification based on a cascade of boosted classifiers is used for the recognition. Experimental results show the efficiency of the proposed algorithm.

1 Introduction

Vision based systems can nowadays be found in many applications, for monitoring goods in private areas or for managing security in public ones. Nevertheless, the relative robustness of vision algorithms, the camera miniaturization and the computation capacity of embedded systems permit other applications of vision based sensors. In order to keep at home low mobility people or to manage the energetic consumption, we need some reliable information on room occupation, the number and the activities of the house occupants. Vision based systems are probably the most efficient technology for this task.

The Capthom project, in which we are working, falls within this context. It consists in developing a human detection low cost sensor. This sensor will present advantages compared to existing sensors, that is to say, a strong immunity to intemperate detections and a great detection reliability. We want to have a reference platform which can give information on a room occupation. This platform will assess the performance of other affordable technologies (e.g. cameras in the visible spectrum). In spite of its prohibitive price, far infrared technology is the most convenient one to automatically watch a room. Acquisition is not influenced by the luminosity. In this framework, we have developed a far infrared algorithm which can detect and track a human in a room.

^{*} This work was realized with the financial help of the Regional Council of Le Centre and the French Industry Ministry within the Capthom project of the Competitive-ness Pole S^2E^2 .

If the need of a reliable human detection system in videos is really important, it is still a challenging task. First, we have to deal with usual object detection difficulties (background variety etc.). Second, there are other specific constraints for human detection. The human body is articulated, its shape changes during the walk. Moreover, human characteristics change from one people to another (skin color, weight etc.). Clothes (color and shape) and occlusions also increase the difficulties.

Some human detection systems have already been proposed in the literature. First, some of them are outline based. They try to recognize the outline which is detected with a foreground segmentation. For instance, Kuno et al. [1] use the outline histogram for the classification. Dedeoglu [2] matches the detected outline with a outline from a database using a distance calculation. Mae et al. [3] compute a distance between the detected outline and a model. These methods are highly influenced by the foreground segmentation step. Moreover, they cannot deal with a moving camera and a crowded scene.

Other methods are based on machine learning. Papageorgiou et al. [4] have first proposed a detector based on Haar wavelets and SVM. Viola et Jones [5] have proposed a global framework for object detection based on the boosting algorithm and Haar wavelets. More recently, Dalal et Triggs [6] have proposed to combine the Histograms of Oriented Gradients with SVM. The good generalization capacity and performance of these techniques are well known. Nevertheless, the learning dataset quality is really important and difficult to set up. Moreover, for video surveillance applications, many available information (e.g. motion and past events) are not used in these methods.

In our project framework, we need a real-time and reliable algorithm. Machine learning algorithms are expensive in computation time but present good performances. So, from this observation, we propose in this article an extension of these machine learning algorithms using advantages given by the video. In our approach, the foreground segmentation is used in order to limit the search space of our classifier. As we don't use the contour of the extracted area for the classification, we are less dependent on the foreground segmentation than outline based methods. Moreover, the 2D tracking system improves the global performance because we have multiple images of the same person at different moments.

Each step of the proposed algorithm (cf. figure 1) will be detailed in this article. First, we apply a foreground segmentation method to localize moving blobs which have a higher temperature than the scene average temperature. Second, after a connected components clustering and a filtering, regions of interest are



Fig. 1. Complete algorithm process

tracked to have chronological information of 2D movements. Then, we try to identify these connected components (that is to say if they are human beings). The performance of the proposed algorithm is shown in the last section with some experimental results.

2 Foreground Segmentation

We choose to first perform a foreground segmentation. The main objective is to simplify the image to proceed without any important information alteration. Only the detected regions of interest will be treated in the following steps. By definition, the background is the union of all static objects and the foreground is composed of all regions in which there is a high probability to find a human. There are two approaches for the foreground segmentation. Some algorithms are based on temporal differencing between two or three successive frames. Other algorithms perform a subtraction between the current frame and a background model. The most used background models are a temporal average, a single Gaussian distribution [7][1], a mixture of Gaussian [9] or a minimum and a maximum [8].

We have shown in [10] that a background subtraction with a single Gaussian distribution presents good performances in terms of detection and computation time. With this model, we take into account the acquisition system noise. As we are working in an indoor environment, we do not have to deal with a multimodal background model. We used this model in our algorithm, computing the average and the variance on each pixel. The foreground detection is realized by the following decision criterion :

$$\begin{cases} B_{1,t}(x, y) = 1 & \text{if } |I_t(x, y) - \mu_t(x, y)| > \tau_1 \cdot \sigma_t(x, y) \\ B_{1,t}(x, y) = 0 & \text{otherwise} \end{cases} \quad (1)$$

where (x, y) are the pixel coordinates and $I_t(x, y)$ its value in gray scale at time t ; $B_{1,t}$ is the binary image of the foreground detection ; μ_t and σ_t are respectively the average and the standard deviation and τ_1 is a threshold set empirically to 2.5.

If $B_{1,t}(x, y) = 0$, the Gaussian model is updated with :

$$\mu_t(x, y) = (1 - \alpha) \cdot \mu_{t-1}(x, y) + \alpha \cdot I_t(x, y) \quad (2)$$

$$\sigma_t^2(x, y) = (1 - \alpha) \cdot \sigma_{t-1}^2(x, y) + \alpha \cdot (I_t(x, y) - \mu_{t-1}(x, y))^2 \quad (3)$$

where α is a threshold determined empirically. Far infrared vision allows to see in night environment and gives also information about the scene temperature. With the hypothesis that the human temperature is noticeably higher than the average of his environment, we perform a binarization to detect hot areas in the image.

$$\begin{cases} B_{2,t}(x, y) = 1 & \text{if } I_t(x, y) > \tau_2 \\ B_{2,t}(x, y) = 0 & \text{otherwise} \end{cases} \quad (4)$$

where $B_{2,t}$ is the binary image representing hot areas, τ_2 is an arbitrary threshold. We finally apply a "logic and" between the background subtraction and the hot area detection result :

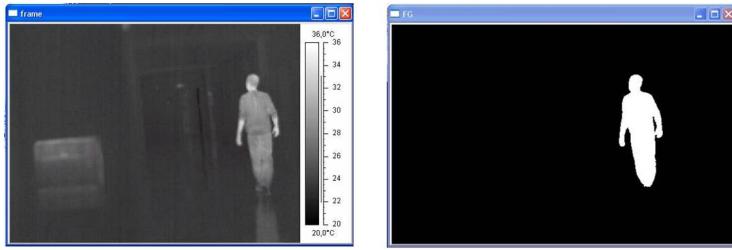


Fig. 2. Foreground segmentation example

$$B_t(x, y) = B_{1,t}(x, y) \cap B_{2,t}(x, y) \quad (5)$$

where B_t is the result of our foreground segmentation. On each frame, we gather detected pixels in connected components, these components are filtered deleting small ones. An example of foreground segmentation result is presented in figure 2.

3 Moving Objects Tracking

Once we have detected and filtered regions of interest in the image, we try to have information about blobs movements. In order to satisfy the real time constraints of our system, we have developed a relatively simple and fast algorithm based on the intersection of the connected components between frames at time t and at time $t - 1$. We compute the matching matrix H_t at time t :

$$H_t = \begin{pmatrix} \beta_{1,1} & \dots & \beta_{1,N} \\ \vdots & \ddots & \vdots \\ \beta_{M,1} & \dots & \beta_{M,N} \end{pmatrix} \quad (6)$$

where M and N are, respectively, the components number of frames at time $t - 1$ and at time t . $\beta_{i,j} = 1$ if the i^{th} component at time $t - 1$ and the j^{th} component at time t intersect, otherwise $\beta_{i,j} = 0$. The analysis of the matrix H_t gives some information on the matching between the components at time $t - 1$ and at time t . For example, if two components a and b at time $t - 1$ and one component at time t intersect we merge components a and b in a new component.

We are able to deal with the merging or the splitting of blobs. However, we do not use any model for our tracked blob and we do not estimate the movement, we are not able to deal with occlusion problems. But, for our application, there are no consequences because, if an object disappears, it is detected and considered as a new object when it appears again.

4 Human Recognition

Once we have chronological information of blobs movements, the following step is to know if the tracked blob is a human. To do that, we build a statistical

model with a machine learning approach. There are many features and learning techniques in the literature. We have chosen the face detection system initially proposed by Viola et Jones [5]. This method is based on Haar wavelets and the boosting algorithm Adaboost.

Our learning dataset is composed of 956 positive images (cf. figure 3) and 3965 negative images (cf. figure 4). Images come from the OTCBVS dataset [12][13] (a benchmark database in the literature) and from images collected with an infrared camera where the ground truths are manually built. Because of the large number of negative images required, we use infrared images but also gray level images in the visible spectrum.



Fig. 3. Example of positive images



Fig. 4. Example of negative images

We use 14 features (Haar-like filters) described in figure 5. Each one is composed of two or three black and white rectangles. The feature values x_i are calculated with a weighted sum of pixels of each component.

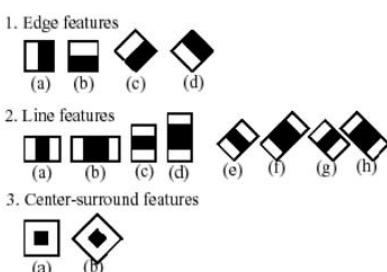


Fig. 5. Haar-like filters

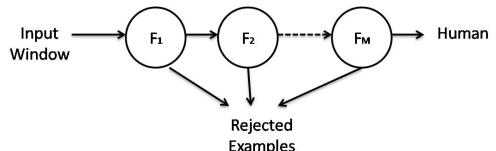


Fig. 6. Cascade of boosted classifiers

Each feature is then used as a weak classifier with :

$$f_i = \begin{cases} +1 & \text{if } x_i \geq \tau_i \\ -1 & \text{if } x_i < \tau_i \end{cases} \quad (7)$$

+1 means it corresponds to a human –1 not. Then, a more robust classifier is built with several weak classifiers using the boosting method [14].

$$F = \text{sign}(c_1 f_1 + c_2 f_2 + \dots + c_n f_n) \quad (8)$$

We build a cascade of boosted classifiers (cf. figure 6). The neighborhood of the connected component tracked in the previous step is successively analyzed by each boosted classifier which can reject or accept the window.

5 Experimental Results

The speed of our detector is closely related with the number and the size of regions of interest. Approximatively, on a standard PC (Windows, 2GHz), for a 564*360 frame size, our algorithm is able to proceed 30 frames per second when there is no blob in the image and 15 to 20 frames per second when there is at least one blob. A detection example is shown figure 7. The ellipse is the result of the blob detection, then we apply our detector in the neighborhood of this blob. A rectangle is drawn around the human if one is detected.

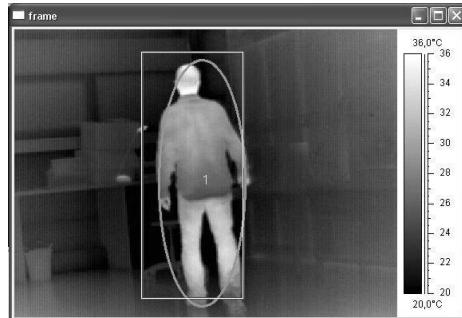


Fig. 7. A detection example, the rectangle is the human detection result in the blob neighborhood shown by the ellipse

We have also evaluated the performance of our classifier. We have built a dataset composed of 317 images. Each image contains at least one human, some infrared images are taken in indoor (room, office) and some in outdoor, some images contain pets.

The figure 8 shows a precision-recall curve obtained on this test dataset where :

$$\text{precision} = \frac{\text{right positives}}{\text{right positives} + \text{false positives}} \quad (9)$$

$$\text{recall} = \frac{\text{right positives}}{\text{right positives} + \text{false negatives}} \quad (10)$$

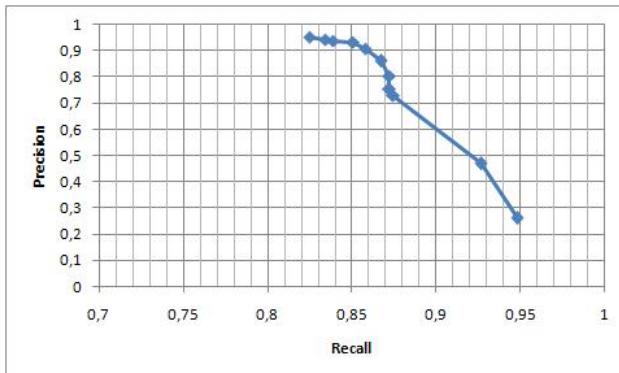


Fig. 8. Precision/Recall curve



Fig. 9. Examples of right positive detections



Fig. 10. Example of false positive and false negative detections

This curve has been obtained by varying the number of stages in the cascade. We obtain up to 95.34% for precision with 20 stages in the cascade and up to 94.78% for recall with 10 stages.

In figure 9 we show some examples of right positive detections. In figure 10 we show one example of false positive and false negative detections. In infrared vision, many surfaces are reflective, this is the cause of the majority of our false

positive detections. Moreover, when someone wears a coat which is at ambient temperature, the distinction with the background is not easy. These two main difficulties are represented in figure 10.

Our classifier has been evaluated on static images but the global performance of our system will be improve with the tracking module.

6 Conclusion and Perspectives

Information on room occupation is really important for many systems, but human detection in video or in images is still a challenging task. In this article, we propose an extension of object detection systems using advantages given by the video. In our approach, the foreground segmentation is used in order to limit the search space of our classifier. Moreover, the 2D tracking system improves the global performance because we have multiple images of the same person at different moments.

Experimental results show the efficiency of our approach. Nevertheless, it remains several ways of improvement. First, the classifier performance is closely related to the database quality, and our database of infrared images can be improved. Second, we have to learn several classifiers for one human. As we work in indoor environment, occlusions are frequent, we could improve the robustness if we learn a part of the body which is more often visible (e.g. head and shoulder). A fusion with the visible spectrum can also decrease the number of false positive detections (because of the reflective surfaces in the infrared spectrum). Finally, we plan to develop our system in order to recover high-level information on human activities in a room.

References

1. Kuno, Y., Watanabe, T., Shimosakoda, Y., Nakagawa, S.: Automated Detection of Human for Visual Surveillance System. In: Proceedings of the International Conference on Pattern Recognition, pp. 865–869 (1996)
2. Dedeoglu, Y.: Moving object detection, tracking and classification for smart video surveillance, PhD thesis, bilkent university (2004)
3. Mae, Y., Sasao, N., Inoue, K., Arai, T.: Person detection by mobile-manipulator for monitoring. In: The Society of Instrument and Control Engineers Annual Conference (2003)
4. Papageorgiou, C., Oren, M., Poggio, T.: A general framework for object detection. In: 6th International Conference on Computer Vision, pp. 555–562 (1998)
5. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the conference on Computer Vision and Pattern Recognition, pp. 511–518 (2001)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 886–893 (2005)
7. Yoon, S., Kim, H.: Real-time multiple people detection using skin color, motion and appearance information. In: Proc. IEEE International Workshop on Robot and Human Interactive Communication, pp. 331–334 (2004)

8. Haritaoglu, I., Harwood, D., David, L.S.: W4: real-time surveillance of people and their activities. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 809–830 (2006)
9. Stauffer, C., Grimson, E.: Adaptive background mixture models for real-time tracking, CVPR, 246–252 (1999)
10. Benerezeth, Y., Emile, B., Rosenberger, C.: Comparative Study on Foreground Detection Algorithms for Human Detection. In: *Proceedings of the Fourth International Conference on Image and Graphics*, pp. 661–666 (2007)
11. Han, J., Bhanu, B.: Detecting moving humans using color and infrared video. In: *Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 228–233 (2003)
12. Davis, J., Keck, M.: A two-stage approach to person detection in thermal imagery. In: *Proc. Workshop on Applications of Computer Vision* (2005)
13. Davis, J., Sharma, V.: Background-Subtraction using Contour-based Fusion of Thermal and Visible Imagery. *Computer Vision and Image Understanding*, 162–182 (2007)
14. Schapire, R.E.: The boosting approach to machine learning: An overview. In: *MSRI Workshop on Nonlinear Estimation and Classification* (2002)

Uniqueness Filtering for Local Feature Descriptors in Urban Building Recognition

Giang Phuong Nguyen and Hans Jørgen Andersen

Department of Media Technology and Engineering Science,
Aalborg University, Denmark
`{gnp,hja}@cvmt.dk`

Abstract. Existing local feature detectors such as Scale Invariant Feature Transform (SIFT) usually produce a large number of features per image. This is a major disadvantage in terms of the speed of search and recognition in a run-time application. Besides, not all detected features are equally important in the search. It is therefore essential to select informative descriptors. In this paper, we propose a new approach to selecting a subset of local feature descriptors. Uniqueness is used as a filtering criterion in selecting informative features. We formalize the notion of uniqueness and show how it can be used for selection purposes. To evaluate our approach, we carried out experiments in urban building recognition domains with different datasets. The results show a significant improvement not only in recognition speed, as a result of using fewer features, but also in the performance of the system with selected features.

1 Introduction

Urban buildings usually have a lot of structural features such as windows, doors, and columns. For this reason local descriptors are widely used in building recognition to describe the content of captured images. A comprehensive overview of existing local descriptors can be found in [1]. Due to their stability under image transformation, partial occlusions and viewpoint changes, interest points are used in creating descriptors. The usual procedure in this approach is first to localize interest points, then to describe a region around each point to form a descriptor of that region. Well-known techniques in this regard are SIFT (Scale Invariant Feature Transforms) [2] and other SIFT-based approaches [3].

While most of the current systems that apply local descriptors employ SIFT in the original or a modified version, we have investigated a more recent technique known as Multi-Scale Oriented Patches (MOPS) that was developed by Brown et al. [4]. In previous work [5] we have analyzed the features of MOPS in detail and compared its performance in building recognition to that of SIFT. In this respect we concluded that MOPS performs better in urban building recognition, especially when large temporal variations are involved. In this paper, we also employ MOPS.

A major issue in approaches that use local detectors is that they usually produce a large number of interest points, i.e. a large set of descriptors is extracted for each image. The number of descriptors greatly depends on the content of the image. For example, SIFT detectors on average create ~ 2000 descriptors for an image of 500x500 pixels [2]. As mentioned at the outset, buildings in general are highly structured, and this means that the number of interest points can go much higher than this. Since the computational cost of matching is positively correlated with the number of descriptors extracted, this problem should be taken into account. Recently, more attention has been paid to recognition speed in application systems, since this is an essential requirement in applications such as robot tracking [6][7], and to run-time recognition with mounted devices [8]. To meet this speed requirement, different approaches have been proposed to improve local detectors. In [3], the authors present a method of reducing the dimensionality of SIFT descriptors, using the PCA dimensional reduction method which projects the original SIFT feature space from 128 dimensions to 20 dimensions. The PCA-SIFT method achieves significant space benefits and requires a third of the time in the matching phase compared to the original SIFT. A different approach is put forward in [9] where a vocabulary tree is used to index descriptors. The K-means algorithm is used to cluster all descriptors and place them in the correct branch. For each query image, extracted descriptors are traced down the tree, a score list is given for all leaves, and the one with the highest score is returned as the best match. This approach has proved to be very fast and scalable to a very large number of descriptors.

The above approaches do not alter the original number of features. However, not all features are equally important. Some detected features are irrelevant in the recognition phase. In such cases, having too many descriptors will reduce the recognition rate. For this reason attention should be focused only on those features that are informative. Examples of techniques used for this purpose can be found in [10][11][14], where the authors concentrate on selecting a subset from given descriptors. For instance, in [10], the authors apply a method designed to select discriminative features best suited for characterizing individual locations in indoor office environments. They produced a high location recognition rate using only 10% of the original detected features. They were therefore able substantially to speed up recognition. In this paper, we focus on developing a new technique for selecting a subset of descriptors that should meet two essential requirements. The first requirement is that of speed. The second is that, while reducing information from the original descriptors, the system should perform as well as or better than existing systems. Throughout the paper, we will use the two terms “features” and “descriptors” interchangeably.

The paper is organized as follows. First, in section [2], we give a brief description of the MOPS detector. This technique is used as the basis for our approach to extracting initial descriptors. In section [3], we present an overview of existing techniques for defining informative descriptors. We then propose our own method of selecting a subset of features. In the next section, we apply our technique to a real application in urban building recognition with two different datasets.

The results are evaluated and discussed in section 4. Finally, we present our conclusions in section 5.

2 Review of the MOPS Detector

Multi-scale oriented patches is a technique recently proposed by Brown et al. for detecting local image features [4]. Let us take an image set \mathcal{I} . Each input image $I_i \in \mathcal{I}$ is incrementally smoothed with a Gaussian kernel $\{\sigma_t\}_{t=1..n}$. An image pyramid is then constructed by down-sampling the image at rate r . In the second step, interest points are extracted using the Harris corners detector at each level of the pyramid. This step yields a set of points at locations where the corner strength is a local maximum of a 3×3 neighborhood and above a threshold of 10 [4]. In the next step, the sub-pixel precision is found by means of a Taylor expansion (up to the quadric term) at those extreme points. Each extreme point is described in terms of its orientation within a window of size 28×28 (corresponding to a Gaussian kernel with $\sigma = 4.5$), and through sampling of grey level values in a 40×40 neighborhood. The grey level values are sampled in a grid with a spacing of 5 pixels rotated according to the orientation. This gives a feature vector for each landmark consisting of 8×8 grey level values. Before matching, the feature vector is standardized by subtracting the mean and dividing it by its standard deviation. Then, as in [4], we perform a Haar wavelet transform on the 8×8 descriptor patch to form a feature vector of 64 dimensions F_j . Finally, for each image we obtain a set of descriptors $\mathcal{F}_{I_i} = \{F_j\}_{j=1..k}$, where k is the number of descriptors extracted from the image I_i .

3 Selecting a Subset of Descriptors

In developing techniques for selecting descriptors, it is generally assumed that certain descriptors are more important than others. The terms “discriminative” and “informative” are usually used to describe significant descriptors. In [10], the authors observe that certain features are more stable and thus better able to handle variations in scale and viewpoint. They therefore aim to select such features. For each feature extracted by means of the SIFT detector from each image at each location, they calculate a posterior probability. The probability values are used as ranking criteria. In [4], the authors present an adaptive non-maximal suppression (ANMS) algorithm that selects a subset of interest points based on their corner strength. The general idea of this algorithm is that for each point extracted through the process described above, they calculate the corner strength, then select points that are maximum within a neighborhood of radius k pixels. In all their experiments, the authors select a maximum of 500 points for each image. This means a set of 500 descriptors is used to describe the content of an image. Another technique for selecting informative (i-SIFT)descriptors,

using the SIFT detector, can be found in [11]. For each given image, informative descriptors are defined as those that appear in discriminative regions. These regions are detected on the basis of an entropy-coded image derived by calculating posterior distribution.

We propose a different way of defining discriminative descriptors. They should identify the most salient features of a given building. One salient property is rarity [12][13]. We therefore define discriminative descriptors as those that are almost unique. Such descriptors maximize discrimination between objects. We thus propose to select descriptors on the basis of their uniqueness i.e. their rarity within a descriptor set.

Definition: A unique feature has an identifiable property that distinguishes it from other features in the image.

In other words, in a feature space where all descriptors are located, the unique feature is that which has the fewest features within its ϵ -neighborhood. Embarking from this definition, we now present our method of selecting unique features. Given an image I , assume that I has k descriptors or feature vectors $\mathcal{F}_I = \{F_1, F_2, \dots, F_k\}$. To calculate the uniqueness of each feature, we first compute the dissimilarity values between feature vectors. For the MOPS descriptors, L2 distance is used as the dissimilarity function. We have $S_{ij} = \sqrt{\sum_{l=1}^t (f_l^i - f_l^j)^2}$, where f_l^i and f_l^j are components of feature vectors F_i and F_j respectively, and $t = 64$. Each feature vector F_i is compared to the others $\{F_1, \dots, F_{i-1}, F_{i+1}, \dots, F_k\}$. We then obtain a set of dissimilarity values $\{S_{i1}, S_{i2}, \dots, S_{i,i-1}, S_{i,i+1}, \dots, S_{ik}\}$. To decide whether two feature vectors are similar or not, an ϵ neighborhood is established. If a feature point F_j in the given feature space falls within the ϵ -neighborhood of F_i , it is considered similar to F_i , i.e.

If $S_{ij} < \epsilon$ then F_i and F_j are similar,
else then F_i and F_j are dissimilar.

As indicated in our definition of a unique feature, we will select features that have the smallest number of similar features within their ϵ -neighborhood. This means that the greater the number of neighbors, the less unique the feature is. Hence, the uniqueness of a feature vector F_i in image I is formulated as:

$$\mathcal{U}_{F_i} = \|\{S_{ij} < \epsilon\}\|_{j=1..k, j \neq i}$$

4 Application in Urban Building Recognition

Building recognition, or object recognition in general, often involves dealing with great variations in image content due to differences of scale, orientation, illumination, and viewpoint. Moreover, different buildings often have very similar structures. Selecting unique features that are able to distinguish one building from another is an essential issue. In this section, we will present our approach to urban building recognition.

4.1 Experimental Setup

We select two different datasets. The first, AAU dataset, contains 135 images captured in the Aalborg University area. The set includes a total of 21 buildings, with an average of 6 images for each building. In this dataset, the buildings are of very similar architectural design. In the second dataset (Aalborg city center), we use a set of 442 images of 19 buildings in the center of Aalborg. The main difference between this dataset and the AAU set is that the images from the city center were taken at different times of day (morning, afternoon, and evening), on different days, and under different weather and seasonal conditions (sunny, cloudy, winter, summer). Moreover, the buildings in the center vary in structure, have more decoration and are more often occluded by passing vehicles or people. These factors together pose a considerable challenge. The following shows more detail on how we created this dataset:

- 16/11/2006: during daytime (a cloudy day).
- 17/11/2006: in the evening with electrical lights on.
- 28/11/2006: during daytime (with Christmas decoration).
- 29/11/2006: during daytime (with Christmas decoration).
- 05/12/2006: in the evening (with Christmas decoration).
- 03/05/2007: sunny day (with Danish flags decoration).

Once the two datasets were created, we applied the MOPS detector to each one. Default parameters of the MOPS were used, and all extracted descriptors stored. Next, we applied our method of selecting unique features of each image. These unique features were stored separately. Since the extraction of unique feature from the images in the datasets was done offline, we did not count the time needed for this step. Instead, we compared our performance with that achieved by other methods. To evaluate performance during the matching process, we reported precision values. Each image in the dataset was sequentially used as a query. The query was then compared to all other images in the corresponding dataset. The top five best matches were returned, and the precision values calculated for each of these. The baseline is the performance achieved by using the default MOPS detector on all extracted descriptors. We also applied the so-called adaptive non-maximal suppression (ANMS)selection method described in [4], which selects the strongest features on the basis of corner strength. For a fair comparison, we performed experiments with different numbers of selected features, namely 100, 200, 300, 500, 800, and 1000. In all experiments, we identified unique features using $\epsilon = 0.3$.

4.2 Experimental Results

First, we present the results from different cases in which varying numbers of descriptors were used, as shown in table 2 and 11. k^* represents our uniqueness filtering method. In these tables, the last row is the default MOPS with all features taken into account. The second columns report the average number of descriptors per image. The third columns report the total descriptors of each

Table 1. AAU dataset with different numbers of unique features per image vs. default MOPS detector

Method	Features/image	Total features
k^*	100	13500
k^*	200	27000
k^*	300	40500
k^*	500	67255
k^*	800	108000
k^*	1000	130831
default	1470	198479

Table 2. Aalborg dataset with different numbers of unique features per image vs. the default MOPS detector

Method	Features/image	Total features
k^*	100	44200
k^*	200	88400
k^*	300	132600
k^*	500	221000
k^*	800	353600
k^*	1000	442000
default	2160	954409

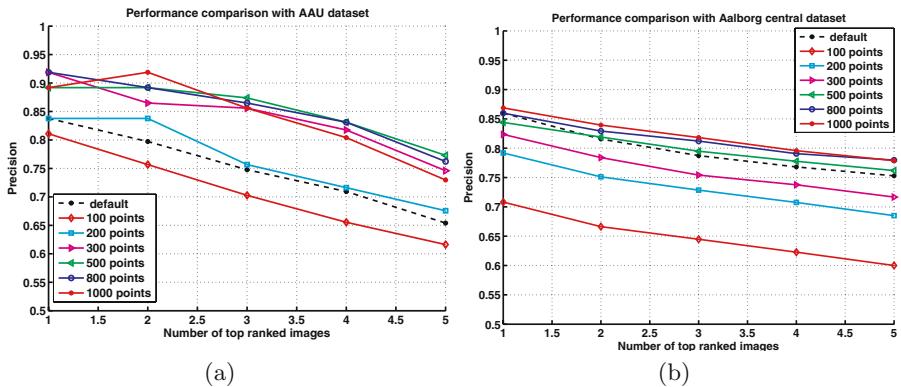


Fig. 1. a. AAU dataset; b. Aalborg city center dataset. Precision vs. the number of top ranked images. Results show performance when the default MOPS with all extracted features is used, versus our approach which focuses only on unique features.

dataset. They show that, on average, the number of features extracted in the default cases is much higher, especially where the Aalborg city center dataset is concerned.

In figures 1a and 1b, we show recognition results for the two datasets. In these figures, we compare performance when different numbers of unique features are used, versus default performance when all extracted features are employed. The figures show that although the default MOPS has the highest number of descriptors, it performs less accurately than our method, in which smaller numbers of descriptors are used. The results achieved by using a certain number of unique features are significantly better, especially in the case of the AAU dataset. With this dataset, even selecting as few as 200 descriptors per image produces better results than the default approach. This means that instead of using all features, we can limit ourselves to only $\sim \frac{1}{7}$ of the total number. As noted in the above section, the Aalborg city center dataset presents a more difficult challenge. Here, at least 300 features were needed to enable reliable recognition of a given

building, although even where fewer features were used we still obtained a rather high recognition rate of 70% in the best match. However, with 500 unique features, i.e. fewer than $\sim \frac{1}{4}$ of the total descriptors, we achieved the same performance as the default approach. Further improvement is shown with 800 and 1000 unique features. We can also see from the two figures that there is a saturation point at which there is little improvement in performance between 800 or 1000 descriptors. This means more descriptors are unnecessary and may even reduce performance by creating disturbance.

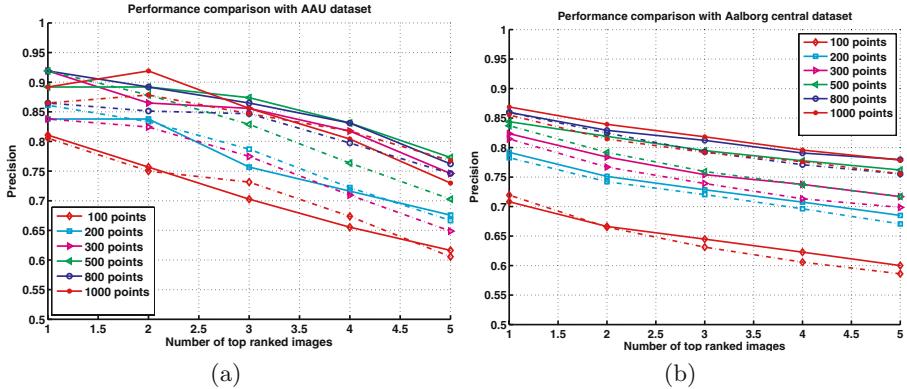


Fig. 2. a. AAU dataset; b. Aalborg city center dataset. Precision vs. the number of top ranked images. Results show performance using our approach vs. ANMS in selection of descriptors.

In the next experiment, we compared the performance achieved by our approach with that achieved when ANMS was used for selecting descriptors. Figures 2a and 2b show the results of the two approaches with different numbers of selected descriptors. The dotted lines represent the results when the ANMS approach was used, and the solid lines the results when our uniqueness method was applied. With 100 or 200 selected descriptors, the performance of the two approaches was comparable in both datasets. Where a higher number of descriptors was used, the results achieved by our method were superior. For example, when 300 descriptors are selected in the AAU dataset, we can achieve a recognition rate of {92%, 87%, 86%, 82%, 75%}, while with ANMS we get {84%, 82%, 77%, 71%, 65%} for the top 1, 2, 3, 4, and 5 matches respectively.

In figure 3, we show some examples of buildings with 500 selected descriptors. The left column shows the results produced with our approach, and the right column the results produced with ANMS. In general, the descriptors selected by the two methods are quite close. ANMS selects features on the basis of their corner strength. Thus the features selected are mainly corners. In our case, we do not consider corners the most important features since they will recur in other similar areas. The features selected should be those that are least like other

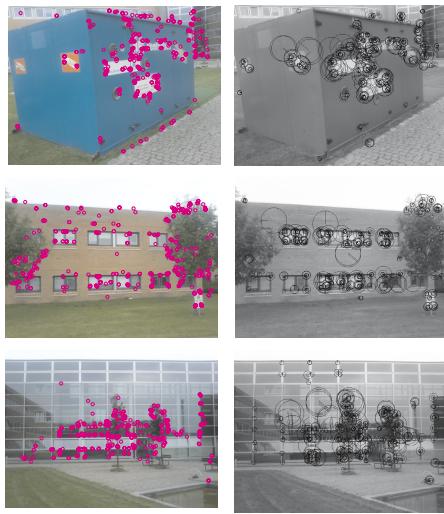


Fig. 3. Examples of buildings with 500 selected descriptors using uniqueness criteria vs. ANMS

features in the same image: i.e. where the smallest number of features in each image resemble them. Since corners are not our first priority in selection, other salient details will be chosen instead.

The proposed approach works well in the applications tested. In general, corner points are good features to detect. When it comes to distinguishing between different buildings, however, corners are not the most discriminative features. This has been shown by our experiments, where the selected points were not always corners. Further, the experiments with both datasets show that the default parameters of the MOPS detector are not optimal in building recognition. It is difficult to determine the number of descriptors that should be extracted per image. With fewer descriptors you certainly get the advantage of greater recognition speed, but on the other hand may not produce sufficient information for matching purposes. The results show that more descriptors are required to capture the content of a complex scene.

5 Conclusion

We have presented a new approach to selecting informative descriptors of an image, based on identifying unique features. Our experimental results with different datasets show that our approach improves recognition performance even where far fewer descriptors are used. Moreover, reducing the number of descriptors also speeds up the matching process, and this is an important factor for any run-time application. However, the time needed for the complex process of extracting all

the descriptors before the selection still remains. Finding informative features without extracting all descriptors is of interest to us in our future work.

Acknowledgement

This research is supported by the IPCity project (FP-2004-IST-4-27571), a EU-funded Sixth Framework program Integrated project on Interaction and Presence in Urban Environments.s

References

1. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 27(10), 1615–1630 (2005)
2. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2004)
3. Ke, Y., Sukthankar, R.: Pca-sift: A more distinctive representation for local image descriptors. In: *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 506–513 (2004)
4. Brown, M., Szeliski, R., Winder, S.: Multi-image matching using multi-scale oriented patches. In: *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, pp. 510–517 (2005)
5. Nguyen, G.P., Andersen, H., Christensen, M.: Urban building recognition during significant temporal variations. In: *IEEE Workshop on Application of Computer Vision* (2008)
6. Royer, E., et al.: Localization in urban environments: monocular vision compared to a differential gps sensor. In: *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 114–121 (2005)
7. Kosecka, J., Li, F., Yang, X.: Global localization and relative positioning based on scale-invariant keypoints. *Robotics and Autonomous Systems* 52(1), 27–38 (2005)
8. Robertson, D., Cipolla, R.: An image based system for urban navigation. In: *British Machine Vision Conference* (2004)
9. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 2161–2168 (2006)
10. Li, F., Kosecka, J.: Probabilistic location recognition using reduced feature set. In: *IEEE Intl. Conf. on Robotics and Automation*, pp. 3405–3410 (2006)
11. Fritz, G., Seifert, C., Paletta, L.: Urban object recognition from informative local features. In: *IEEE Intl. Conf. on Robotics and Automation*, pp. 132–138 (2005)
12. Kadir, T., Brady, M.: Saliency, scale and image description. *International Journal of Computer Vision* 2(45), 83–105 (2001)
13. Schiele, B., Crowley, J.L.: Probabilistic object recognition using multidimensional receptive field histograms. In: *International Conference on Pattern Recognition* (1996)

Dynamic Human Fatigue Detection Using Feature-Level Fusion

Xiao Fan, Bao-Cai Yin, and Yan-Feng Sun

Beijing Key Laboratory of Multimedia and Intelligent Software, College of Computer Science and Technology, Beijing University of Technology, Beijing 100022, China
wolonghongni@email.bjut.edu.cn

Abstract. Driver fatigue is a significant factor in many traffic accidents. We propose a novel dynamic features using feature-level fusion for driver fatigue detection from facial image sequences. First, Gabor filters are employed to extract multi-scale and multi-orientation features from each image, which are then merged according to a fusion rule to produce a single feature. To account for the temporal aspect of human fatigue, the fused image sequence is divided into dynamic units, and a histogram of each dynamic unit is computed and concatenated as dynamic features. Finally a statistical learning algorithm is applied to extract the most discriminative features and construct a strong classifier for fatigue detection. The test data contains 600 image sequences from thirty people. Experimental results show the validity of the proposed approach, and the correct rate is much better than the baselines.

Keywords: Computer vision, human fatigue, Gabor filters, fusion, AdaBoost.

1 Introduction

Driver fatigue is a significant factor in many traffic accidents. In China, driver fatigue resulted in 3056 deaths in vehicular accidents in 2004, and caused 925 deaths in highway accidents which amounted to about 14.8%. When a driver fatigues, many special visual features will appear on his face and body. These features reflect a person's level of fatigue. Many computer vision based approaches have been proposed for fatigue detection in the last decade.

The frequency and time of eye closed all increase when driver fatigue. Much attention is paid to eye features for fatigue detection. Based on the study by the Federal Highway Administration, percentage of eyelid closure (PERCLOS) has been found to be the most reliable and valid measure of a person's alertness level among many drowsiness detection measures. Wang *et al.* [1] used Gabor filters to extract texture features of eyes, and used Neural Network classifier to identify drivers' fatigue behavior. The doze stage was judged when the area of the iris becomes below a threshold in [2]. Dong *et al.* [3] decided whether the driver was fatigue by detecting the distance of eyelids.

Yawning is also an important character of fatigue. Wang *et al.* [4] took the mouth region's geometric features to make up an eigenvector as the input of a BP ANN, and they acquired the BP ANN output of three different mouth states that represent

normal, yawning or talking state respectively. Wang *et al.* [5] represented the openness of the mouth by the ratio of mouth height to width, and detected yawning if the ratio was above 0.5 in more than 20 frames.

Most of these methods are spatial approaches. Although spatial approaches can achieve good recognition in some cases, they do not model the dynamics of fatigue and therefore do not utilize all of the information available in image sequences. In facial expression recognition, according to psychologists [6], analyzing an image sequence produces more accurate and robust facial expression recognition. Psychological researches suggest that facial motion is fundamental to facial expression recognition. Therefore, more attention [7, 8] has been shifted particularly towards modeling dynamic facial expressions.

Fatigue is a cognitive state that is developed over time. It is our belief that dynamic features which capture the temporal pattern should be the optimal features to describe fatigue just as facial expression recognition. To account for the temporal aspect of human fatigue, Ji *et al.* [9] introduced a probabilistic framework based on dynamic Bayesian networks for modeling and real-time inferring human fatigue. The dynamic fatigue model integrates fatigue evidences spatially and temporally. However, in summary, there is little research in extracting dynamic features from image sequences for fatigue detection. Achieving high accuracy in fatigue detection is still a challenge due to the complexity and variation of facial dynamics.

In this paper, to account for the temporal characteristic of human fatigue, we propose a novel dynamic features using feature-level fusion to detect fatigue from image sequences. Fig.1 illustrates the framework of the proposed approach. First, after each image in a face image sequence is preprocessed by face detection, geometric normalization and cropping, Gabor filters are used to extract multi-scale and multi-orientation features from each image in the sequence. Then these features are merged according to a fusion rule to produce a single feature. To get the dynamic features of human fatigue, the fused image sequence is divided into rectangle region sequences as dynamic units, and a histogram of each dynamic unit is computed and concatenated as dynamic features. Finally, weak classifiers are constructed on the histogram features and AdaBoost is applied to select the most discriminating dynamic features and build a strong classifier for fatigue detection.

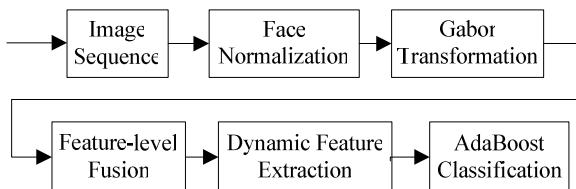


Fig. 1. Framework of the proposed approach

The remainder of this paper is organized as follows. Section 2 introduces Gabor transformation on image sequences. Dynamic feature extraction is showed in section 3. In section 4, feature selection and classifier learning are described. Finally, experimental results and conclusions are presented in section 5 and 6.

2 Multi-scale and Multi-orientation Representation

We exploit Gabor filters to decompose the input face image sequence and get the multi-scale and multi-orientation representation for the sequence after original face images are preprocessed and normalized in a predefined way. Face images in our experiments are normalized with a size of 64*64.

2.1 Gabor Filters

Gabor filters [10] share many properties with the receptive fields of simple cells in the mammalian visual cortex. Due to their biological relevance and computational properties, they have been successfully used in a variety of image processing applications such as face recognition. Gabor filters can be defined as follows:

$$\varphi_j(\vec{x}) = k_j^2 \exp\left(-\frac{k_j^2 x^2}{2\sigma^2}\right) \exp\left(i k_j \vec{x}\right)$$

Where, $\vec{k}_j = \begin{pmatrix} k_{jx} \\ k_{jy} \end{pmatrix} = \begin{pmatrix} k_v \cos\phi_u \\ k_v \sin\phi_u \end{pmatrix}$, $k_v = 2^{-\frac{v+2}{2}} \pi$, $\phi_u = u \frac{\pi}{K}$, and v determines the frequencies of the Gabor filters, u determines the orientations of the Gabor filters. For an image with the gray level distribution of $L(\vec{x})$, at a given point $\vec{x} = (x, y)$, The Gabor representation of the image is the convolution of the image with a family of Gabor filters defined as follows:

$$J_j(\vec{x}) = \int L(\vec{x}') \varphi_j(\vec{x} - \vec{x}') d^2 \vec{x}'$$

In our experiments, we adopt a discrete set of Gabor filters which comprise 5 spatial frequencies, i.e., scales, and 8 distinct orientations. We set $K=8$, $v \in \{0, 1, 2, 3, 4\}$, $u \in \{0, 1, 2, 3, 4, 5, 6, 7\}$.

2.2 Representation for an Image Sequence

The representation for a face image sequence can be derived by convoluting it with the multi-scale and multi-orientation Gabor filters. Each image in the sequence is convolved with the 40 Gabor filters (5 scales and 8 orientations) to generate the Gabor features. Thus, for each pixel position in the face image, 40 complex values can be calculated as Gabor features.

Given one image sequence I with n images, we label each image with I_i , where i is the index of the image. We label the Gabor features of image I_i with $G_{i,u,v}(x, y)$, where i is the index of the image and v is the index of the Gabor filter scale, u is corresponding to the Gabor filter orientation. Based on each image $G_{i,u,v}(x, y)$, we obtain a multi-scale and multi-orientation representation G as $\{G_{i,u,v}(x, y) : i \in (0, \dots, n-1), u \in (0, \dots, 7), v \in (0, \dots, 4)\}$ for the image sequence I . Fig. 2 illustrates the representation (magnitudes of the Gabor features) of a face image.

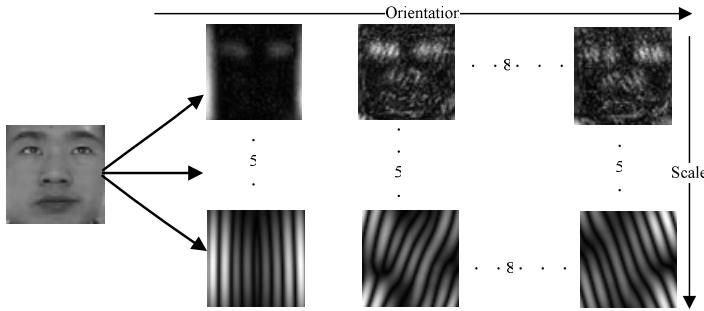


Fig. 2. Multi-scale and multi-orientation representation of a face image

3 Dynamic Feature Extraction Based on Feature-Level Fusion

To obtain the dynamic features of human fatigue, multi-orientation features are merged according to a fusion rule to produce a single feature, and the fused image sequence is divided into region sequences as dynamic units. Then, a histogram of each dynamic unit is computed and concatenated as dynamic features.

3.1 Feature-Level Fusion

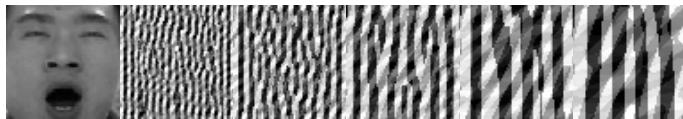


Fig. 3. Original image and multi-scale fused images (real part)

In order to extract the local orientation information, the multi-orientation features are fused into a single code. Gabor filters that we adopted include 8 distinct orientations. The local orientation is estimated by thresholding the 8-orientation features of each pixel into binary codes and considering the result as a binary number. We threshold Gabor features to $\{0, 1\}$ codes by the following formula:

$$T_{i,u,v}(x, y) = \begin{cases} 1 & \text{if } G_{i,u,v}(x, y) > 0 \\ 0 & \text{if } G_{i,u,v}(x, y) \leq 0 \end{cases}$$

Where, $G_{i,u,v}(x, y) u \in (0, \dots, 7)$ corresponds to the values of 8-orientation Gabor features at the pixel (x, y) . The resulting 8 codes are considered as a binary number. The decimal form of the fused code can be expressed as follows:

$$C_{i,v}(x, y) = \sum_{u=0}^7 T_{i,u,v}(x, y) * 2^u$$

There are a total of 256 possible values and each code value represents a type of local orientation. Decimal forms of the fused codes are computed at each scale. Finally, we get 5-scale fused images for one image (Fig.3).

3.2 Dynamic Feature Extraction

To obtain the dynamic features, the fused image sequence is further divided into small non-overlapping rectangle region sequences to enhance its shape information, from which the local histograms are extracted and concatenated into a single extended histogram to capture the temporal information (Fig. 4).

In the experiments, we divide each fused image into 8×8 regions. Each rectangle region can be denoted as $R_{i,v,r}(x, y)$ ($i \in (0, \dots, n-1)$, $v \in (0, \dots, 4)$, $r \in (0, \dots, 63)$). The same regions in the fused image sequence are concatenated into a dynamic unit (region sequence) $S_{v,r}(x, y)$ as $\{R_{0,v,r}(x, y), R_{1,v,r}(x, y), \dots, R_{n-1,v,r}(x, y)\}$. A histogram of a dynamic unit can be defined as:

$$h_{i,v,r} = \sum_{x,y} I\{S_{v,r}(x, y) = i\}, i = 0, \dots, 255, \text{ where } I\{A\} = \begin{cases} 1, & A \text{ is true} \\ 0, & A \text{ is false} \end{cases}$$

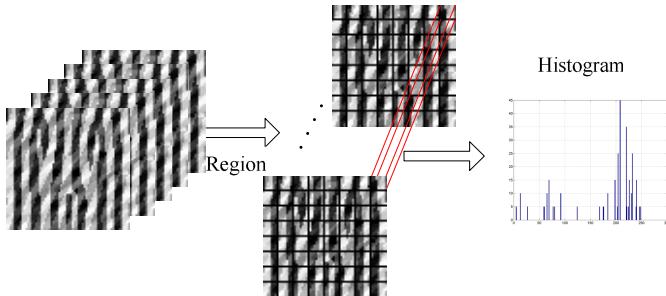


Fig. 4. Flowchart of dynamic feature extraction

There are 256 bins for one dynamic unit. Each histogram bin is the number of occurrences of the corresponding code in a dynamic unit. The resulting histograms on each dynamic unit are combined yielding the extended histogram for an image sequence. The histogram to describe an image sequence can be defined as

$$H = \{h_{i,v,r} : i \in (0, \dots, 255), v \in (0, \dots, 4), r \in (0, \dots, 63)\}$$

With this histogram, we effectively have a dynamic description of a face image sequence on spatial and temporal levels.

4 Statistical Learning of Best Features and Classifiers

The number of dynamic features of an image sequence is $256 \times 64 \times 5 = 81920$, which is too high dimensional for fast extraction and accurate classification. We use AdaBoost both to select a small set of features and train the classifier.

4.1 Weak Classifiers

The weak classifiers are the basis of AdaBoost. In our case, the weak classifiers are decision trees based on the histogram features. Decision tree is a tree graph, with

leaves representing the classification results and nodes representing some predicates. Branches of the tree are marked with true or false.

4.2 AdaBoost Learning

AdaBoost [8, 11] can select features and build a strong classifier at the same time. It provides a simple yet effective stage-wise learning approach for feature selection and nonlinear classification. The final hypothesis of AdaBoost algorithm is

$$H(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^T \alpha_i h_i(x) \geq \frac{1}{2} \sum_{i=1}^T \alpha_i \\ 0 & \text{otherwise} \end{cases}$$

The strong classifier is a linear combination of the T weak classifiers. The AdaBoost learning procedure is aimed to derive α_i and $h_i(x)$.

5 Experimental Results

5.1 Database

The test sets in a lot of researches are often very small with only several subjects. To test the fatigue detection methods, we built a fatigue face database. We used web cameras to catch videos of about forty persons. The videos of each person last several hours which are caught indoors without directions to the subjects.



Fig. 5. Examples of (a) normal and (b) fatigue image sequences in the database

Totally, we got about 50GB videos in AVI format compressed by MPEG-4. We selected thirty subjects' face fatigue videos from the original ones. Then we extracted the fatigue image sequences from the videos and made up the fatigue face database. Totally there are 600 image sequences of 10 female subjects and 20 male subjects. Each subject has 10 normal image sequences and 10 fatigue image sequences. There are 5 images with a resolution of 320*240 in each image sequence (Fig. 5).

5.2 Baselines

We present two statistical learning methods, one based on LDA (PCA+LDA classifier) and one on HMM (PCA+HMM classifier) as baselines. PCA+HMM employs

HMM to classify the sequence of PCA coefficients extracted from a face image sequence. PCA+LDA employs LDA to classify the PCA coefficients of face images. Two experiments were made on the LDA based classifier. In the first one, LDA is used to classify the PCA coefficients from a single image. In the other, PCA coefficients from an image sequence are combined into a feature vector and classified by LDA.

5.3 Results

In our experiments, we implemented Real AdaBoost scheme. The proposed approach was applied to the Gabor real and imaginary parts. The best performance with decision trees of different numbers of splits are showed in Fig. 6. We can see that the proposed approach gets the highest correct rate when the decision trees have one split. Then, when the number of splits increases, the performance decreases.

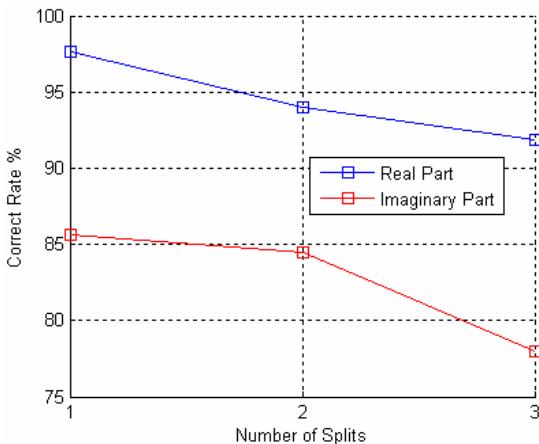


Fig. 6. Best performances with different numbers of splits

Table 1. Confusion matrix for fatigue detection

Method		Normal	Fatigue
PCA+HMM	Normal	74.33%	25.67%
	Fatigue	23.33%	76.67%
PCA+LDA (Single)	Normal	78.27%	21.73%
	Fatigue	18.73%	81.27%
PCA+LDA (Sequence)	Normal	82.00%	18.00%
	Fatigue	17.67%	82.33%
Our Method (Real)	Normal	97.33%	2.67%
	Fatigue	2.00%	98.00%
Our Method (Imaginary)	Normal	79.67%	20.33%
	Fatigue	8.33%	91.67%

The recognition results of all the classifiers are listed in table 1. The average correct rates are 75.50%, 79.77%, 82.17%, 97.67% (real), and 85.67% (imaginary) respectively. The proposed approach with the Gabor real part achieves a very encouraging recognition rate which is much better than the other methods. The correct rate of normal sequences is 97.33% while that of fatigue is 98.00%. PCA+LDA (sequence) is better than PCA+LDA (single) in performance. This may be account for the dynamic characteristic in the sequence. The performance of PCA+HMM is poor. The reason may be that HMM needs enough training samples.

6 Conclusions

A novel dynamic feature using feature-level fusion is presented to account for the spatial and temporal aspects of human fatigue in image sequences. AdaBoost algorithm is used to extract the most discriminative features and construct a strong classifier for fatigue detection. The proposed approach is validated in a real-life fatigue environment with 600 image sequences from thirty people. Experiment results show the validity of the proposed approach, and an encouraging correct rate is achieved. In addition, this approach can be easily extended to video based facial expression recognition. Future efforts will be focused on how to combine the dynamic features from Gabor real parts and imaginary parts to get better performance.

Acknowledgments. This research is sponsored by National Natural Science Foundation of China (No.60533030), Beijing Natural Science Foundation (No.4061001), and PHR (IHLB).

References

1. Wang, R., Guo, K., Shi, S., Chu, J.: A Monitoring Method of Driver Fatigue Behavior Based on Machine Vision. In: Proceedings on intelligent Vehicles Symposium, pp. 110–113 (2003)
2. Miyakawa, T., Takano, H., Nakamura, K.: Development of Non-contact Real-time Blink Detection System for Doze Alarm. In: SICE, vol. 2, pp. 1626–1631 (2004)
3. Dong, W., Wu, X.: Fatigue Detection Based on the Distance of Eyelid. In: IEEE International Workshop on VLSI Design and Video Technology, pp. 365–368 (2005)
4. Wang, R., Guo, L., Tong, B., Jin, L.: Monitoring Mouth Movement for Driver Fatigue or Distraction with One Camera. In: The 7th International IEEE Conference on Intelligent Transportation Systems, pp. 314–319 (2004)
5. Wang, T., Shi, P.: Yawning Detection for Determining Driver Drowsiness. In: IEEE International Workshop on VLSI Design and Video Technology, pp. 373–376 (2005)
6. Bassili, J.: Emotion Recognition: The Role of Facial Movement and the Relative Importance of Upper and Lower Areas of the Face. J. Journal of Personality and Social Psychology 37, 2049–2059 (1979)
7. Zhao, G., Pietikainen, M.: Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. J. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(6), 915–928 (2007)

8. Yang, P., Liu, Q., Metaxas, D.N.: Boosting Coded Dynamic Features for Facial Action Units and Facial Expression Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–6 (2007)
9. Ji, Q., Lan, P., Looney, C.: A Probabilistic Framework for Modeling and Real-time Monitoring Human Fatigue. J. IEEE Transactions on Systems, Man and Cybernetics, Part A 36(5), 862–875 (2006)
10. Wiskott, L., Fellous, J.M., Kuiger, N., von der Malsburg, C.: Face Recognition by Elastic Bunch Graph Matching. J. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 775–779 (1997)
11. Freund, Y., Schapire, R.E.: A Decision-theoretic Generalization of On-line Learning and an Application to Boosting. J. Journal of Computer and System Sciences 55(1), 119–139 (1997)

Real Time Monocular Depth from Defocus

Jean-Vincent Leroy¹, Thierry Simon², and François Deschenes^{1,3}

¹ Centre de MOIVRE, Université de Sherbrooke, Sherbrooke (Québec) Canada J1K2R1

² Université de Toulouse ; UTM-IUT de Figeac ; LRP-mip (Laboratoire de recherche pluridisciplinaire du nord-est de Midi-Pyrénées) ; Avenue de Nayrac, F-46100 FIGEAC, France

³ Université du Québec en Outaouais, Gatineau (Québec) Canada J8X3X7

Jean-vincent.leroy@usherbrooke.ca, thierry.simon@univ-tlse2.fr, françois.deschenes@usherbrooke.ca

Abstract. The method proposed in this paper uses two blurred images, acquired from the same point of view with different focus settings, in order to estimate depth in real time. It falls under the group "Depth from defocus" methods. The blur is modelled by the convolution of a Gaussian Point Spread Function (PSF) which with the theoretical sharp image. We calculate the gradients and Laplacians of the two blurred images and according to the type of contour, step, ramp, roof or line, we compute the differences of these images or their derivative until order 2 to obtain the difference in blur, difference of the variances of the Gaussian ones, which we connect to the depth by taking account of the parameters of the optical system. We use then this difference to the depth by taking account of the parameters of the optical system. We present a set of results on real images showing the performance of the method and its limits. The computing times which we measured make possible the use of our algorithm at video rate.

Keywords: Depth from defocus, image processing.

1 Introduction

We propose a new method to estimate depth of a scene observed by a single camera. The three-dimensional information lost by projection onto the image plane can be get back using a second image. This acquisition is done from the same point of view with different focus settings. This modification leads to variations of optical blur between the two images. The proposed method estimates defocus blur at pixel locations that correspond to intensity discontinuities, that is to say edges. It thus belongs to the category of DFD (Depth from Defocus) methods [1-4].

Our work is the continuation of a previous method proposed by Simon et al. which is based on the use a blurred image and a sharp one [5]. The difficulty to obtain a sharp image led Simon et al. to extend their method in order to use three blurred images [6]. However the acquisition of three images increases both complexity of the process and computing time.

The depth map obtained remains however sparse matrix. A group of solution using the textured areas of the image makes it possible on the contrary to obtain a dense map for example [8]. These solutions model a zone of the image by a set of functions by exploiting the successive derivatives until order n. In these methods the computing times become prohibitory for applications where the execution at video rate is necessary.

To address those issues, we propose in this paper a new and simple method that is able to compute depth from defocus at video rate using two blurred images of the same scene. This method is based on a simple use of the Laplacian magnitude of the two images.

The technique that our propose uses on the theoretical elements developed initially by Subbarao [2] then generalized by Deschênes and Al in [3] exploiting the optical blur on textured zones. Our algorithm determines the type of edges and constitutes the difference of the images or their successive derivatives to obtain the difference in blur which we know to connect to the depth. This method thus makes it possible to obtain the depth for scenes of which the image present of the zones textured or not. The textured zones are regarded as very dense zones of contours, discontinuities of luminance.

We chose to apply the methodology developed by Deschênes and al. [3] approach by textured areas, to discontinuities of luminance marking contours as well as textures.

Section 2 briefly presents the theoretical background behind our DFD method. Section 3 presents our new method. Section 4 is dedicated to experimental results with real images. Finally section 5 presents the limits of our method and the perspectives of development.

2 Theoretical Background

2.1 Relationship between Blur and Depth

The figure 1 presents an ideal optical system in the case of a thin lens of focal length f , and an aperture of diameter L . For a given object point located at a distance s_o of the lens plane corresponds a single point (i.e., sharp image) at a distance s_i . Any movement of the sensor plane and/or the object along the optical axis causes blur, that is an object point now corresponds to a circular disk on the image plane, called circle of confusion, whose size is estimate by its diameter $2R$.

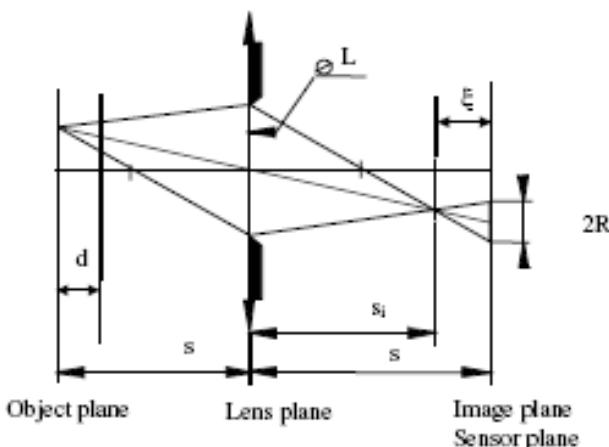


Fig. 1. Blurred image formation process for given focal length and aperture

The Gauss law of the thin lenses:

$$\frac{1}{s_o} + \frac{1}{s_i} = \frac{1}{f}$$

makes it possible to establish an explicit relationship between the diameter of the circle of confusion and the parameters of the optical system:

$$2R = L s_i \left(\frac{1}{f} - \frac{1}{s_o} - \frac{1}{s_i} \right)$$

For $\xi \geq 0$, the sensor plane is behind the focal plane, otherwise it is in front of the focal plane. From optical relationships, the distance to an object (commonly called depth) can be expressed by the following equations:

$$s_o = \frac{fs_i}{s_i - f - 2R \frac{f}{L}} \quad \text{for } \xi < 0$$

$$s_o = \frac{fs_i}{s_i - f + 2R \frac{f}{L}} \quad \text{for } \xi \geq 0$$

These expressions are the basis of all methods that use the optical blur to compute depth.

2.2 Image Formation Process: Modeling Blur

A 2D image point of a given 3D object point can be considered as the sum of the contributions of points belonging to the surface in the vicinity of this 3D point that reflect light toward the camera. This process can be modeled by a 2D convolution [7] between an ideal image $I_s(x, y)$, that is which is sharp every where, and the point spread function (PSF), $h(x, y, s_o)$, of the optical system.

$$I_b(x, y, s_o) = I_s(x, y) \otimes_{2D} h(x, y, s_o)$$

The PSF is obviously a function of the camera parameters but is also function of the depth. It is commonly modeled by the Pillbox function or the Gaussian function [9]. In what follows, we will consider the PSF is Gaussian:

$$h(x, y, s_o) = G(\sigma_{s_o}) = \frac{1}{2\pi\sigma_{s_o}^2} \exp\left(-\frac{x^2 + y^2}{2\sigma_{s_o}^2}\right)$$

The diameter of the circle of confusion thus depends on the scale parameter σ_{s_o} of the Gaussian function. The relationship between those two parameters is given by $\sigma_{s_o} = k2R$, where k is a proportionality constant.

The camera parameters f, L, k, s_i can first be determined by calibration. Hence, being able to estimate σ_{s_o} , the remaining unknown, would allow us to estimate the depth s_o .

2.3 Edge Model

In order to process video in real time, we focus on a subset of features of the image: the edges. An edge is a discontinuity of luminance.

Specifically, we consider three edge models, $C_1(x, y)$, $C_2(x, y)$, and $C_3(x, y)$, that are respectively a step edge, a ramp and roof. For instance, in the case of vertical edges, C_1, C_2, C_3 can be represented as follows:

$$C_1(x, y) = \begin{cases} a & x < x_o, \forall y \\ a + b & x \geq x_o, \forall y \end{cases}$$

which is a function of Heaviside of amplitude (b-a) and of offset a, and

$$C_2(x, y) = \begin{cases} a & x < x_o - \frac{\varepsilon}{2}, \forall y \\ a + \frac{(b-a)}{\varepsilon} \left(x - x_o + \frac{\varepsilon}{2} \right) & x_o - \frac{\varepsilon}{2} \leq x \leq x_o + \frac{\varepsilon}{2}, \forall y \\ a + b & x > x_o + \frac{\varepsilon}{2}, \forall y \end{cases}$$

which is a ramp function of amplitude (b-a), of width ε and inflexion point x_0 .

$$C_3(x, y) = \begin{cases} a & x < x_o - \varepsilon_1, \forall y \\ a + \frac{(e-a)}{\varepsilon_1} (x - x_o + \varepsilon_1) & x_o - \varepsilon_1 \leq x \leq x_o, \forall y \\ b + \frac{(e-b)}{\varepsilon_2} (x_o + \varepsilon_2 - x) & x_o \leq x \leq x_o + \varepsilon_2, \forall y \\ b & x > x_o + \varepsilon_2, \forall y \end{cases}$$

Which is roof or line model, roof for large value of ε_1 et ε_2 .

For edges that are not vertical the previous two models must simply be applied in the direction of the gradient. Particular attention must be paid to the sampling step which highly depends on the orientation. In order to make independent from the orientation one may use interpolation techniques.

We will apply a specific detection according to the type of edges previously definite starting from an analysis of the magnitude of the gradients of the image.

For example, for edges of the C_1 type, the thresholded magnitudes of gradients give the pixels of contour directly. For contours of the type C_2 or C_3 we use the zero-crossing of the Laplacian.

3 Our Method

We propose to locally model blurred edges by development in series of functions, of many authors [2][3] showed that the difference of the two blurred images can then be expressed by:

$$I_{b1}(x, y) - I_{b2}(x, y) = \frac{1}{2}(\sigma_2^2 - \sigma_1^2)\nabla^2 I_b$$

where

$$\nabla^2 I_b(x, y) = \frac{\nabla^2 I_{b1}(x, y) + \nabla^2 I_{b2}(x, y)}{2}$$

and $I_{b1}(x, y) = I_s(x, y) \otimes G(\sigma_1)$, $I_{b2}(x, y) = I_s(x, y) \otimes G(\sigma_2)$ and ∇^2 Laplacian operator.

We can write : $I_{b2}(x, y) = G(\beta) \otimes I_{b1}(x, y)$, with $\beta^2 = (\sigma_2^2 - \sigma_1^2)$, $\sigma_2 > \sigma_1$.

In the same way the partial derivatives are dependent according to the same type of relation:

$$I_{b1}^{(i)(j)} - I_{b2}^{(i)(j)} = \frac{1}{2}(\sigma_2^2 - \sigma_1^2)\nabla^2 I_b^{(i)(j)} \text{ with } I_b^{(i)(j)} = \frac{\partial^{i+j} I_b}{\partial x^i \partial y^j}.$$

Moreover authors showed that the difference in blur β^2 makes it possible to calculate the depth by [8] :

$$s_0^{-1} = \frac{1}{f} - \frac{1}{s_i + s} \left(1 + \sqrt{1 + \frac{(s_i + s)k^2 \beta^2}{L^2(s - s_i)}} \right) \quad (1)$$

The parameters of the optical system are evaluated in a phase of calibration.

These relations usually used on textured zones, in our new method, are applied to edges. For a better estimate of β^2 we actually calculate:

$$\beta^2 = 2 \frac{t_{(0)(0)} \psi |D^{(0)(0)}| + t_{(1)(0)} |D^{(1)(0)}| + t_{(0)(1)} |D^{(0)(1)}|}{t_{(0)(0)} \psi |\nabla^2 I_b^{(0)(0)}| + t_{(1)(0)} |\nabla^2 I_b^{(1)(0)}| + t_{(0)(1)} |\nabla^2 I_b^{(0)(1)}|}$$

where $D^{(i)(j)} = I_{b2}^{(i)(j)} - I_{b1}^{(i)(j)}$ and if $|\nabla^2 I_{b1}^{(i)(j)}| < L_{\min}$ or $|\nabla^2 I_{b2}^{(i)(j)}| < L_{\min}$ or

$|\nabla^2 I_b^{(i)(j)}| < L_{\min}$ or $\beta_{(i)(j)}^2 < \beta_{\min}$ or $\beta_{(i)(j)}^2 > \beta_{\max}$ then $t_{(i)(j)} = 0$ else $t_{(i)(j)} = 1$, with

$\beta_{(i)(j)}^2 = \frac{\nabla^2 I_b^{(i)(j)}}{2(I_{b1}^{(i)(j)} - I_{b2}^{(i)(j)})}$. Actually we compute the value of β^2 on a small surface of the image of size $c \times d$:

$$\beta^2 = \frac{2 \sum_c \sum_d t_{(0)(0)} \psi |D^{(0)(0)}| + t_{(1)(0)} |D^{(1)(0)}| + t_{(0)(1)} |D^{(0)(1)}|}{\sum_c \sum_d t_{(0)(0)} \psi |\nabla^2 I_b^{(0)(0)}| + t_{(1)(0)} |\nabla^2 I_b^{(1)(0)}| + t_{(0)(1)} |\nabla^2 I_b^{(0)(1)}|} \quad (2)$$

The derivatives are calculated by the application of the operator of Prewitt, and the Laplacian by the successive application, twice, of $\sqrt{A_x^2 + A_y^2}$ with A_x and A_y the two directional masks of Prewitt.

The following algorithm summarizes the method:

1. Computing of the directional derivatives by the operators of Prewitt.
2. Computing of Laplacian.
3. Detection of edges on the derivative of the least blurred image
4. For each point of edge:
 - a. Computing β^2 by (2) on area with the size $c \times d$.
 - b. Computing the depth by (1).

4 Results

The results which we present here make it possible to judge the effectiveness of the method. Figure 2 presents the real image, least blurred, with three plans in-depth, of which two outsides are slowly inclined and with various textures. The depth varies between 790 and 990mm, the image has a resolution of 1700 X 1000 pixels.

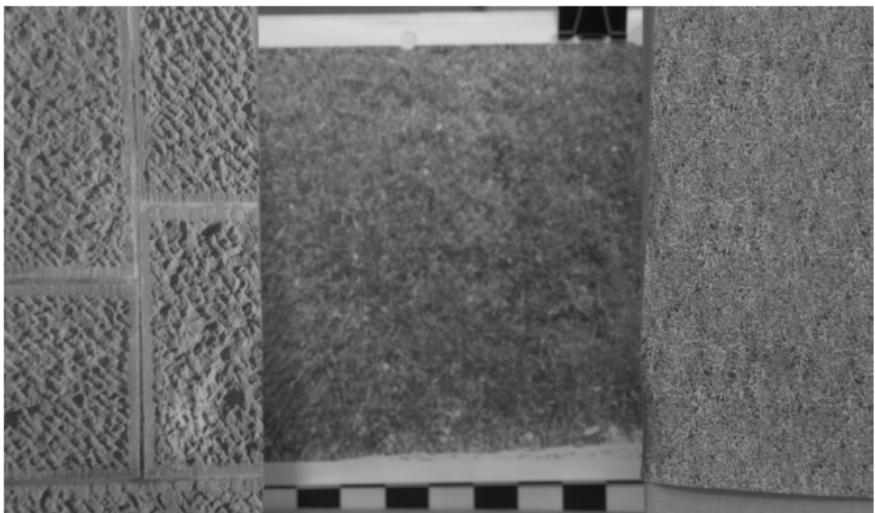


Fig. 2. The least blurred image

Figure 3 gives the results of the computing of β .

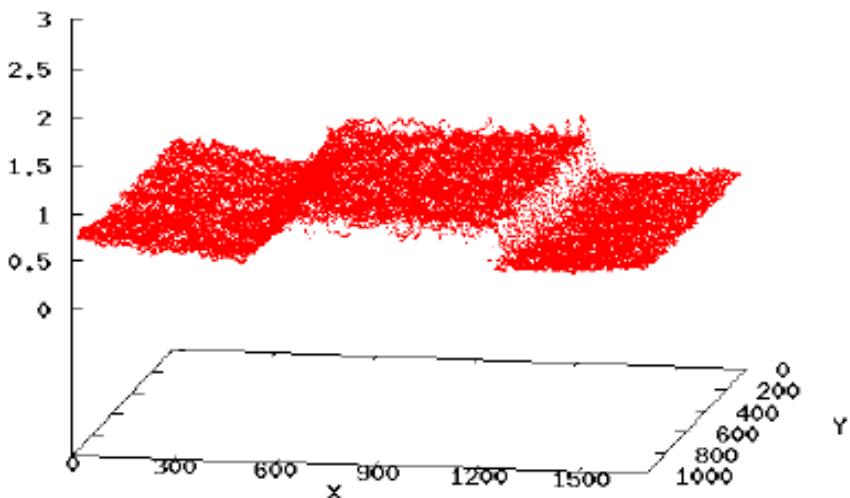


Fig. 3. β Estimations

Figure 4 shows the results obtained for the depth. The computing time is of 80ms what corresponds to approximately 23ms for a size of 800x600 pixels, traditional size for the standard cameras. This computing time makes it possible to consider a processing at the video rate. The mean errors are respectively of 11.03mm, 20.05mm and 5.44mm for the plan of right-hand side, the medium and left. It appears that according to the density of edges, characteristic of textures, the results vary. They are better when the density is high.

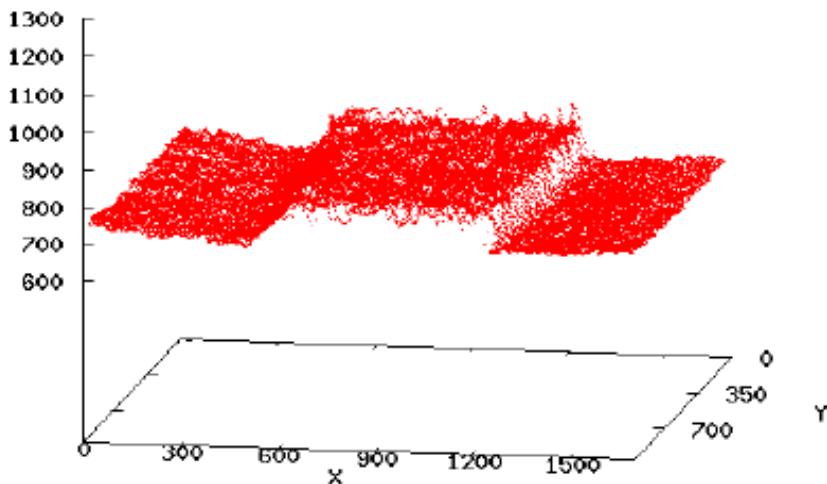


Fig. 4. Depth Estimations

Computations were carried out on APPLE Mac Pro bi-processors, Bi-hearts INTEL Xeon, at 3 GHz. Figure 5 shows a little textured scene to which we apply our method. As we could envisage it the results, figure 6, are worse, taking into account the absence of texture. The depths remain within the limits of real dimensions but they are less stable.

To carry out calculations the value $\psi = 3$ gave us the best results and the values L_{\min} , β_{\min} , β_{\max} were given in experiments. They seem independent of the type of images; we respectively fixed them at 0.5, 0.1, 6. Finally the zone of computing of the mean was selected to 7 X 7.

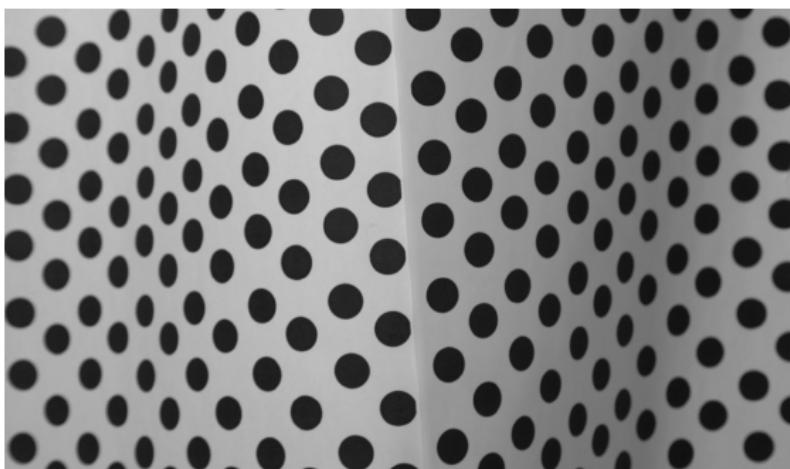


Fig. 5. The least blurred image

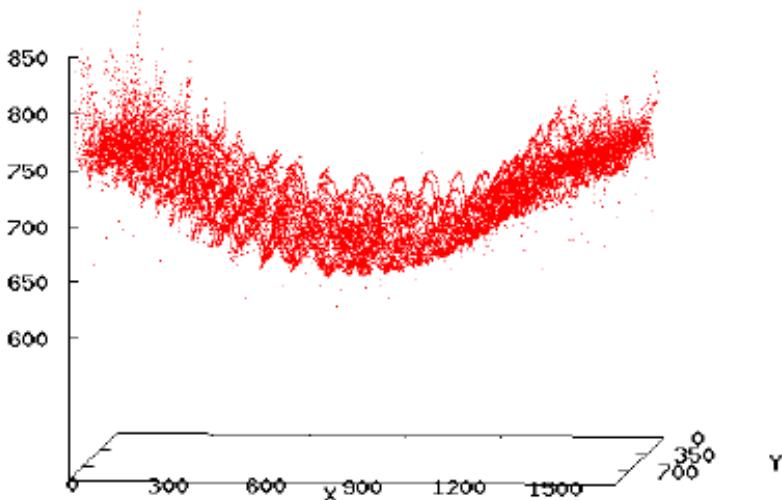


Fig. 6. Depth Estimations

5 Conclusion

We proposed in this paper a new method to estimate depth of a scene from a single camera. The three-dimensional information lost by projection on the plane sensor is obtained using defocus blur at pixel locations that correspond to edges. This method is based on the computation of the ratio of the laplacian magnitude of the images. It is thus fast. Experimental results confirm that this method lead to accurate results at video rate. In further work, we will increase the accuracy to detect in a first step the discontinuity of the depth.

Indeed the mean calculated on an area of size 7 X 7 makes it possible to limit the influence of the noise but led in the areas of strong variation of depth to a smoothing.

References

1. Lin, H.Y., Chang, C.H.: Depth from motion and defocus blur. Optical engineering 12(45), (127201)- (1-12201-10) SPIE (2006)
2. Subbarao, M., Surya, G.: Depth from defocus: a spatial domain approach. Int. J. Comp. Vision 13, 271–294 (1994)
3. Deschênes, F., Ziou, D., Fuchs, P.: An unified approach for simultaneous and cooperative estimatiom of defocus and spatial shifts. Image and Vision Computing n° 22, 35–57 (2004)
4. Favaro, P., Soatto, S.: 3D shape estimation and image restoration: exploiting defocus and motion blur. Springer, London (2007)
5. Simon, C., Bicking, F., Simon, T.: Depth Estimation Based on Thick Edges in Images. In: IEEE ISIE 2004, Ajaccio, France, May 3-8 (2004)
6. Simon, T., Simon, C.: Depth Perception from three blurred images. International Electronic Congress
7. Hopkins, H.H.: The frequency response of a defocused optical system. In: Proc. Of the Royal Society of London series A, pp. 91–103 (February 1955)
8. Ziou, D., Deschênes, F.: Depth from Defocus Estimation in Spatial Domain. Computer Vision and Image Understanding 81, 143–165 (2001)
9. Pentland, A.P.: A new sense for depth of field. IEEE Trans. On Pattern Analysis and Machine Intelligence 9(4), 523–531 (1987)

Dynamic Skin Detection in Color Images for Sign Language Recognition

Michał Kawulok

Institute of Computer Science, Silesian University of Technology, Akademicka 16,
44-100 Gliwice, Poland
michal.kawulok@polsl.pl

Abstract. Skin detection is the first step of processing in many approaches to face and gesture recognition. This paper presents research aimed at detecting skin in digital images for Polish Sign Language recognition. There are many methods for detecting human skin, including parametric skin models defined in various color spaces and statistical approaches which require appropriate training. The presented method is based on statistical model updated dynamically for every image in which human faces can be detected. The detection is performed in luminance channel based on geometric properties of human faces. The experiments proved that effectiveness of this approach is higher than application of general skin detection models.

1 Introduction

Sign languages were developed in communities of deaf people and they are the main means of communication among them. Information is conveyed by gestures and lips movement instead of sound as in the case of natural spoken languages. Noteworthy, the sign languages developed independently in many communities and there is no universal one. Similarly as in the case of spoken languages, the sign languages vary from one geographic region to another, however the coverage of their usage is often not identical to that of spoken languages. For example, the British Sign Language is very different from the American Sign Language (ASL), which is used not only in the area of the US, but also in Mexico and Canada. The research presented in this paper is the first step of a project focused on Polish Sign Language which is used in Poland.

The main problem that appears nowadays is that the majority of people cannot use and do not understand sign languages which creates a barrier separating deaf communities from the rest of society. This barrier can be reduced or even eliminated by automatic systems which translate between spoken and sign languages. A crucial part of these systems is detection and recognition of gestures in digital images. The research presented in this paper is focused on detecting skin areas in order to determine position of faces and hands.

Skin detection is a widely investigated area and many methods for solving this problem have been developed [1, 2, 4, 5, 6, 8, 10, 13, 14]. They can be divided into parametric and statistical methods which are described in section 2. Some

of these methods are quite effective, but they are highly sensitive to changes of lighting conditions and they also produce high false positive errors in case of objects which color is similar to human skin.

Face detection has many applications ranging from entertainment to surveillance tracking and it is also a preliminary stage of face recognition [15]. There are approaches to face detection, which utilize color information to identify skin regions at first and then to build facial feature maps [6]. For frontal face images the detection can be successfully performed in luminance channel based on characteristic properties of face geometry.

The goal of this research was to detect faces and hands which are crucial in case of sign language recognition. At first, faces are detected based on luminance channel. It is feasible, because they are usually oriented frontally to the camera in the analyzed images. After that skin areas are detected using statistical skin color model in *RGB* color space, which is updated dynamically based on detected face region. The experimental results presented in section 5 confirmed that the developed approach is more effective than other skin detectors.

2 Color Based Skin Detection

It was observed [6,13] that skin-tone color has common properties and that it can be defined in various color spaces after applying color normalization. In general, skin detection methods are based on parametric and statistical skin modeling. The main problem concerned with skin color models is their sensitiveness to changes in lighting conditions and high false positive rates in case of objects which color is similar to human skin.

2.1 Parametric Skin Models

Parametric skin models [4,6,10,11,14] are based on fixed decision rules defined empirically in various color spaces after analysis of skin-tone distribution. Those rules are applied to determine if a pixel color value belongs to skin or not. In this research four parametric models in different color spaces were investigated.

Kovac et al. [10] proposed a model defined in *RGB* color space. Skin-tone color has been also modeled in *HSV* space by Nikolaidis [14]. These both models specify conditions for pixel color values and if they are met, a pixel is classified as skin. The models are applicable for various illumination conditions and do not require any color normalization.

An approach proposed by Hsu et al. [6] takes advantage of common skin color properties in nonlinearly transformed YC_bC_r color space. At first, the top 5% of the luminance values in the image define a *reference white* color and the pixel values are scaled linearly in *RGB* space, so that all reference-white pixels have values equal to 255. After the normalization, skin is detected based on an elliptical model defined in nonlinearly transformed YC_bC_r space.

There are also techniques which operate in two color spaces to increase detection accuracy. A composed skin color detector proposed in [11] is defined in *RGB* and YC_bC_r color spaces.

2.2 Statistical Skin Modeling

Parametric methods are aimed at finding models of skin-tone color in a given color space. It is also possible to analyze distribution of skin pixel values for a given training set of images, in which skin and non-skin areas are marked. This allows to create a statistical model, which determines probability that a given pixel value belongs to the skin class. Several methods were developed [2, 5, 8] which implement such approach to skin detection.

At first, based on the training set, histograms for skin (C_S) and non-skin (C_{NS}) classes are built. For every color value (v) probability of observing it in each class is calculated:

$$P(v|C_x) = C_x(v)/N_x , \quad (1)$$

where $C_x(v)$ is number of v -color pixels in x -th class and N_x is total number of pixels in that class. Maximal number of histogram bins depends on pixel bit-depth and usually it is $256 \times 256 \times 256$. However, the experiments showed that detection is more effective if number of bins is reduced to about 64 per channel.

To detect skin with low false positive error it is not only important that a pixel value appears frequently as skin, but also its density in the non-skin class should be low. Hence, probability that a given pixel value belongs to the skin class ($P(C_S|v)$) is calculated based on Bayes rule:

$$P(C_S|v) = \frac{P(v|C_S)P(C_S)}{P(v|C_S)P(C_S) + P(v|C_{NS})P(C_{NS})} , \quad (2)$$

where $P(C_S)$ and $P(C_{NS})$ are estimated based on number of pixels in both classes. If the training set is big enough, probabilities for all possible color values can be estimated to create skin color probability map. However, if a value is not observed in the training set, it is assumed that probabilities $P(v|C_S)$ and $P(v|C_{NS})$ equal 0.5. After obtaining the probability map, every pixel value can be transformed into probability that the pixel belongs to the skin class. Skin regions are extracted based on acceptance threshold.

3 Face Detection

Face detection [6, 15] is aimed at finding location of human faces in digital images. If faces are oriented frontally, which can be assumed in case of sign language recognition, their position is unambiguously defined by central points of eyes. There are many techniques of face detection which utilize common geometrical features of human faces or properties of skin color.

In the presented research face detection was performed in luminance channel based on geometric properties of human faces. The first stage is a preliminary selection of *face candidates*, which are those areas in images where faces may be located. After that, the candidates are verified and those of them which are accepted are regarded as faces (Fig. 1). This approach is effective and faces are detected with high precision [9].

Usually human heads have shape of vertically oriented ellipses, so at first such ellipses are detected in images using Hough transform [7]. After that, small horizontally oriented ellipses are detected inside upper half of every ellipse to find candidates for eye sockets. The eye-socket candidates are verified and if there are two positive responses inside an ellipse, a new face candidate is created.

The candidates are verified with the Support Vector Machines (SVM) [3] which are robust learning machines solving two-class classification problems. In order to use the SVM as a verifier, it is trained with two classes of samples: normalized face images of constant size in which eyes are in invariant positions and normalized non-face candidates. Similarly, the SVM is trained and used for eye sockets verification.

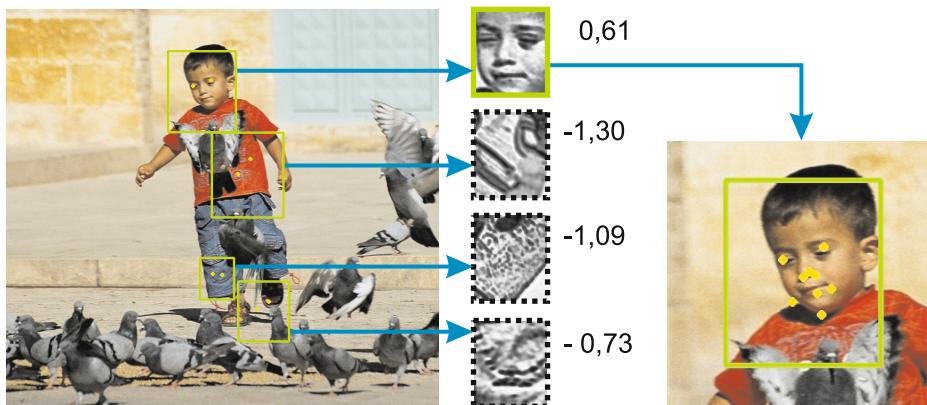


Fig. 1. Example of face detection. Face candidates are detected (*left*), normalized and verified with the SVM (*middle*). If its response (*shown for every normalized image*) is positive, a candidate is regarded as a face.

4 Face-Based Dynamic Skin Model

Skin detection algorithms are vulnerable to lighting condition changes and they generate high false positive errors in case of objects which color is similar to human skin. In vast majority of cases lighting conditions are constant within an image and skin color do not vary much for a single person. Therefore, if skin color could be defined separately for every person, the errors should be significantly decreased. This can be achieved if it is known that a given region of the image contains skin pixels. Based on that region a statistical model can be built or conformed dynamically and applied to the whole image.

If a face is detected in an image, it is possible to point a region which with high probability contains only skin pixels. In the proposed approach such face region is defined based on detected positions of eyes as presented in Fig. 2.

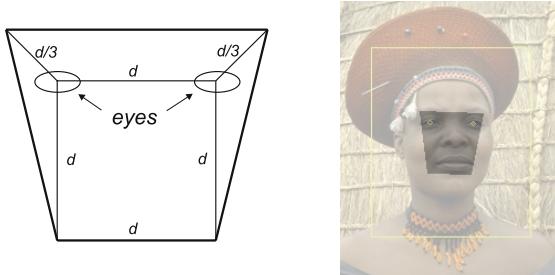


Fig. 2. Face region generation based on positions of eyes

Pixels inside the face region are regarded as skin pixels and probability of observing a given value within the region ($P(v|C_S)$) is calculated [1]. The distribution of non-skin pixel values is unknown and it was estimated that number of non-skin pixels for a given value is:

$$C_{NS}(v) = \max_i(C_S(i)) - C_S(v) , \quad (3)$$

which makes it possible to build the dynamic skin-color probability map [2] based on detected face region. As mentioned in [2.2], optimal number of histogram bins for a general model is 64 per channel. In the presented case only a part of color domain is covered by pixels from the face region and it occurred to be more effective to create the face-based probability map for 16 bins per channel.

A model created in this way takes into account only distribution of skin pixel values, which allows for detecting skin areas effectively, but in some cases false positive error is high. The results occurred to be much better if the face-based model was joined with a general statistical model trained with a large training set as described in section 2.2. Both models are defined by probability maps (P_{face} and P_{gen}) which are merged to obtain the final probability map:

$$P(C_S|v) = \omega P_{face}(C_S|v) + (1 - \omega)P_{gen}(C_S|v) , \quad (4)$$

where ω is a weight of the dynamic model. Based on experiments its value was set to 0.8. Probability maps must have the same pixel value domain, so the face-based map is converted from 16 to 64 bins per channel. If no face is detected in an image, the general model is applied to detect skin regions. Example of skin detection with the described approach is presented in Fig. 3. The background color is similar to human skin which results in very high false positive rate (68.9%) for a general model. False positives were reduced to only 1.1% when the model was merged with the dynamic face-based skin model.

5 Experiments

The experiments were conducted for ECU face detection database [12] which contains 4000 images with marked areas of human skin. The images were divided into two groups: a testing set, for which the detection performance was

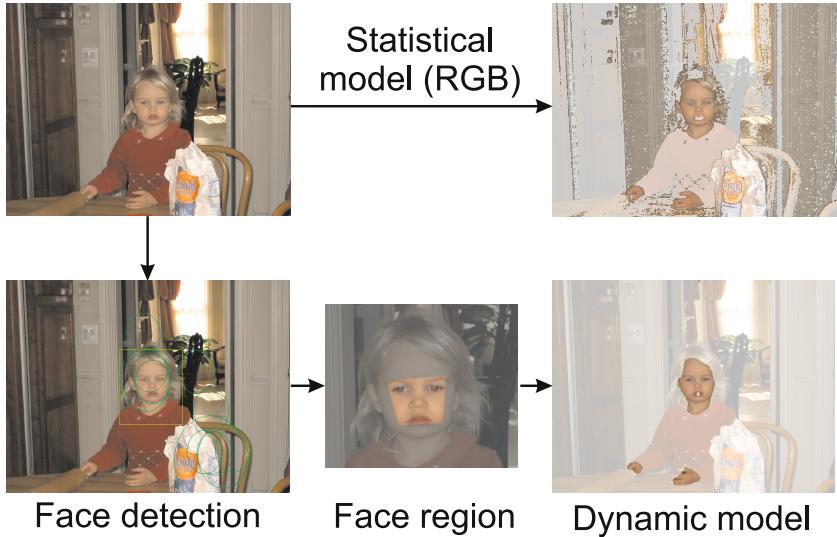


Fig. 3. Skin detection with a statistical model and with a face-based dynamic model

measured, and a training set. Skin detectors were implemented and tested in Adaptive Vision Studio (available at www.adaptive-vision.com). Skin detection effectiveness is measured by: false positive rate (δ_{FP}), which is a percentage of non-skin pixels classified as skin pixels, and false negative rate (δ_{FN}), which is a percentage of skin pixels classified as non-skin. Sum of these rates is treated as detection error used for comparison. In case of statistical models the rates depend on acceptance threshold.

During the experiments several skin detectors were tested: *RGB* detector [10], *HSV* detector [14], a detector in nonlinearly transformed YC_bC_r space [6], composed detector [11], statistical skin model in *RGB*, *HSV* and YC_bC_r and the face-based model developed in the current research. The results are compared in Fig. 4. Performance of statistical models was better than for parametric methods, but the best results were achieved with the face-based model joined with statistical *RGB*. For statistical models acceptance threshold was set to a value, for which sum of the errors was minimal. Choice of a color space for the statistical model do not influence the results significantly. Number of histogram bins per channel is quite relevant, as shown in Fig. 5b for the *RGB* statistical model.

Table 1. False positive error for various false negative rates

Skin detection model	$\delta_{FN} = 5\%$	$\delta_{FN} = 10\%$	$\delta_{FN} = 20\%$
Statistical <i>RGB</i>	23.8%	15.7%	8.8%
Face-based	26.1%	12.6%	5.4%
Face-based with stat. <i>RGB</i>	15.3%	5.1%	1.9%

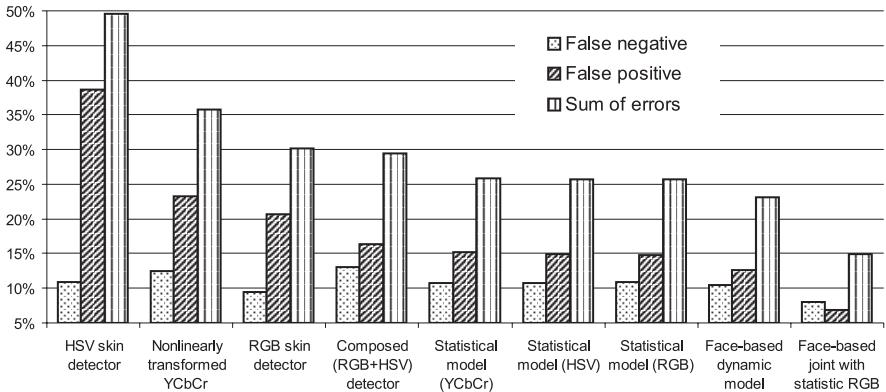


Fig. 4. Comparison of major skin detection methods

Results for *RGB* statistical model, face-based dynamic model and joined model are presented in Fig. 5a in a form of receiver operating characteristic (ROC) curves. Exact values of false positive errors for three false negative rates are presented in Tab. II.

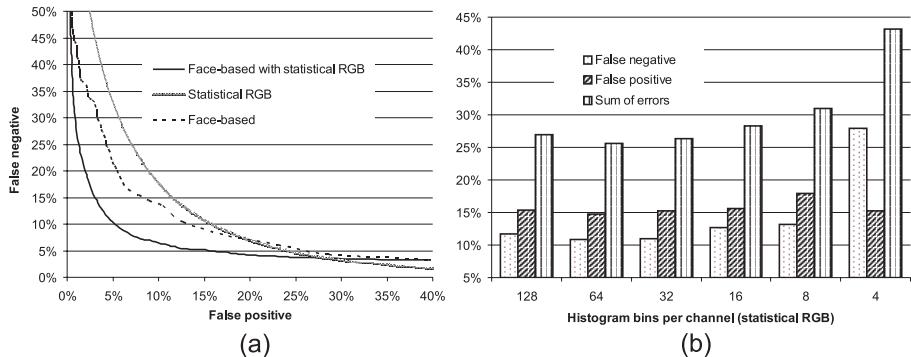


Fig. 5. ROC curves for *RGB* statistical model and face-based models (a). Performance of *RGB* statistical model for various number of histogram bins per channel (b).

6 Conclusions and Future Work

In this paper a concept of dynamic skin detection was presented. Based on face detection performed with Hough transform and the SVM verifier, a statistical skin model is build for every detected face region. After that a general statistical skin model is updated with the face-based model in *RGB* color space. The

developed method was compared with other known skin detectors and it allowed to reduce false positive detection to only 5.1% at 10% false negative rate for the tested skin database.

Future works will be aimed at taking into account not only color information of pixels, but also their neighborhood. It is also planned to detect position of hands in the areas indicated as human skin which will be the next step towards the sign language recognition system.

References

1. Albiol, A., Torres, L., Delp, E.J.: Optimum color spaces for skin detection. In: Proc. of the Int. Conf. on Image Proc., vol. 1, pp. 122–124 (2001)
2. Brand, J., Mason, J.: A Comparative Assessment of Three Approaches to Pixel-level Human Skin-Detection. In: Proc. of the Int. Conf. on Pattern Recognition, vol. 1, pp. 1056–1059 (2000)
3. Cortes, C., Vapnik, V.: Support-Vector Networks. Machine Learning 20(3), 273–297 (1995)
4. Filipe, T., Tiago, C., Hamid, S.: Improved Automatic Skin Detection in Color Images. In: VIIITH digital image comp.: techniques & applications, pp. 419–427 (December 2003)
5. Gomez, G., Moralez, E.: Automatic feature construction and a simple rule induction algorithm for skin detection. In: Proc. of the ICML Workshop on Machine Learning in Comp. Vision, pp. 31–38 (2002)
6. Hsu, R.-L., Abdel-Mottaleb, M., Jain, A.K.: Face detection in color images. IEEE Trans. on Pattern Analysis and Machine Intelligence 24(5), 696–706 (2002)
7. Illingworth, J., Kittler, J.: A survey of the Hough transform. Comp. Vision Graphics and Image Proc. 44(1), 87–116 (1988)
8. Jones, M.J., Rehg, J.M.: Statistical color models with application to skin detection. International Journal of Comp. Vision 46(1), 81–96 (2002)
9. Kawulok, M.: Application of Support Vector Machines in automatic human face recognition. Medical Informatics & Technologies 9, 143–150 (2005)
10. Kovac, J., Peer, P., Solina, F.: Human Skin Colour Clustering for Face Detection. In: Eurocon 2003 Ljubljana, pp. 144–148. Slovenia (2003)
11. Kukharev, G., Nowosielski, A.: Fast and Efficient Algorithm For Face Detection in Colour Images. Machine Graphics & Vision 13(4), 377–399 (2004)
12. Phung, S.L.: ECU Face Detection Database, University of Wollongong (2002)
13. Terrillon, J.-C., David, M., Akamatsu, S.: Automatic Detection of Human Faces in Natural Scene Images by Use of a Skin Color Model and of Invariant Moments. In: Proc. of the 3rd International Conf. on Automatic Face and Gesture Recognition, Nara, Japan, pp. 112–117 (1998)
14. Tsekridou, S., Pitas, I.: Facial feature extraction in frontal views using biometric analogies. In: Proc. of EUSIPCO 1998, vol. 1, pp. 315–318 (September 1998)
15. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. Tech. Report CARTR-948, Center for Automation Research, University of Maryland, College Park (2000)

Depth Estimation by Finding Best Focused Points Using Line Fitting

Aamir Saeed Malik and Tae-Sun Choi

Gwangju Institute of Science and Technology,
1 Oryong-Dong, Buk-Gu, Gwangju, 500712, Korea
{aamir_tschoi}@gist.ac.kr

Abstract. This paper presents a method for depth estimation using image focus based on the linear regression model. For each pixel, we select two datasets based on the maximum value which is calculated using Laplacian operator. Then linear regression model is used to find lines that approximate these datasets. The best fit lines are found using least squares method. After approximating the two lines, their intersection point is calculated and weights are assigned to calculate the new value for the depth map. The proposed method is compared with four depth estimation algorithms.

Keywords: Depth Map, 3D shape recovery, Shape from focus, Line fitting.

1 Introduction

There are many methods for the calculation of depth leading to 3D shape recovery. In this paper, we limit our discussion to one of such methods, i.e., Shape From Focus (SFF). The objective of shape from focus is to find out the depth of every point of the object from the camera lens. Hence, finally we get a depth map which contains the depth of all points of the object from the camera lens where they are best focused or in other words, where they show maximum sharpness.

The basic problem of imaging systems, such as the eye or a video-camera, is that depth information is lost while projecting a 3D scene onto 2D image plane. Therefore, one fundamental problem in computer vision is the reconstruction of a geometric object from one or several observations. Shape information that is obtained from the reconstruction of a geometric object is of critical importance in many higher level vision applications like mobile robot systems. For example, an unmanned spacecraft, in order to land safely on lunar surface, needs to estimate depth details of the terrain. Various image processing techniques retrieve the lost cue and shape information from the pictorial information. Shape from focus (SFF) is one of such image processing techniques that are used to recover such information.

Various techniques and algorithms have been proposed in the literature for the implementation of SFF. They include methods using focus image surface, Lagrange polynomial, neural networks, dynamic programming etc. But almost all the techniques start with the estimation of the depth map. Hence, the techniques for the estimation of this initial depth map become quite significant.

Generally SFF scheme relies on a Focus Measure operator and an approximation technique. Focus Measure operator plays a very important role for three dimensional shape recovery because it is the first step in calculation of the depth map. So a focus measure operator needs to show robustness even in the presence of noise. Hence it should provide a very good estimate of the depth map.

2 Related Work

2.1 Focus Measure Methods

A Focus Measure operator is one that calculates the best focused point in the image. And focus measure is defined as a quantity to evaluate the sharpness of a pixel locally. Franz Stephan Helml and Stefan Scherer [1] summarized the traditional focus measures while introducing three new focus measure operators.

Laplacian, the most commonly used operator, is suitable for accurate shape recovery because of being a symmetric operator, and is obtained by adding second derivatives in the x and y directions. Modified Laplacian (ML) [2,3] is computed by adding squared 2nd derivates. In order to handle possible variations, Shree K. Nayar and Yasuo Nakagawa suggested a variable spacing (step) between the pixels used to compute derivatives. In order to improve robustness for weak-texture images, Shree K. Nayar and Yasuo Nakagawa [2] presented focus measure at (x,y) as sum of ML values in a local window (about 5x5).

Tenenbaum Focus Measure is gradient magnitude maximization method that measures the sum of squared responses of horizontal and vertical Sobel masks. Variance Focus Measure is based on the variance of gray-level which is higher than that in a blur image. Mean Method Focus Measure [1,3] depends on the ratio of mean grey value to the center grey value in the neighborhood. The ratio of one shows a constant grey-level or absence of texture. Curvature Focus Measure [1] exploits that the curvature in a sharp image is expected to be higher than that in a blur image. Point Focus Measure [1] is approximated by a polynomial of degree four.

2.2 Approximation Methods

A more accurate depth range image can be obtained by using some optimization and approximation method. The results of the focus measures, defined in section 2.1, are refined using such a reconstruction scheme. First we discuss the traditional SFF method. In Traditional (TR) SFF, for each image in the sequence, focus measure at each pixel is computed by Sum Modified Laplacian in a 2D neighborhood around the pixel. The results of TR SFF are improved by Subbarao and Choi [4] who proposed a new concept termed Focused Image Surface (FIS) based on planar surface approximations. The FIS of an object is defined as the surface formed by the set of points at which the object points are focused by a camera lens. Aamir Malik and Tae-Sun Choi [5] summarized various approximation techniques.

FIS can be improved by a piecewise curved surface rather than piecewise planar approximation. This was proposed by Choi and J. Yun [6]. They estimated the piecewise curved surface by interpolation using second order Lagrange polynomial. Asif and Choi [7] used Neural Networks to learn shape of FIS by optimizing the focus

measure over small 3-D windows. Bilal and Choi [8] proposed the use of Dynamic Programming (DP) to handle the complexity of FIS. DP is motivated by the Bellman's principal of optimality. A direct application of DP on a 3D data is impractical due computational complexity. Therefore, a heuristic model based on DP was proposed by Bilal and Choi.

3 Method

An initial estimate of depth map is obtained using the SML focus measure. For every pixel, its maximum value is obtained by considering the entire sequence of image frames. A set of successive points are selected leading upto the maximum value (with maximum value being the last point in the set). Then we use linear regression to find a line that approximates this data set. Again another set of successive points are selected following the maximum value (with maximum value being the first point in the set). Procedure is repeated to find a line that approximates this data set. After approximating the two lines, their intersection point is calculated. The new value for the depth map is calculated by assigning weights based on the distance of intersection point from the maximum value and the distance of intersection point from the closest value in the two data sets. An averaging filter is used to minimize noise effects. We also used optimum window size 5x5 [9].

Let $I(x,y,k)$ be the acquired image sequence where k represents the total number of frames. $I_m(x,y,k)$ is acquired by processing each of these frames using sum of modified Laplacian method [2,3]. Let P be any pixel (x,y) then P_{\max} is calculated as:

$$P_{\max}(x, y) = \max[I_m(x, y, k)], \text{ where } k = 1, \dots, n \quad (1)$$

The value of k that gives P_{\max} is referred to as k_{\max} . Now, two datasets (DS_1 and DS_2) are generated with one leading upto P_{\max} while the other following P_{\max} .

$$DS_1(t) = P(x, y, k_{\max} - t); \text{ where } t = 0, 1, \dots, j \text{ & } j < k_{\max} \quad (2)$$

$$DS_2(t) = P(x, y, k_{\max} + t); \text{ where } t = 0, 1, \dots, j \text{ & } j < k_{\max} \quad (3)$$

$$\text{when } t = 0, DS_1 = DS_2 = P_{\max}$$

In this paper, we used $t=0,1,\dots,4$. We found that $j=4$ results in optimum performance for our algorithm. After generating the two datasets, we used the least square linear regression for the data sets to approximate two lines. The method of finding the line that best fits the collected data is known as linear regression while the least-squares regression line minimizes the sum of the squares of the residuals. The best fit line associated with the n points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ has the form:

$$y = mx + b \quad (4)$$

$$\text{slope : } m = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$\text{Intercept : } b = \frac{\sum y - m \sum x}{n}$$

The best fit line is calculated using the least squares method. After calculating the best fit lines for the two datasets (DS_1 and DS_2), the intersection point of the two lines is found. Let P_i be the point of intersection of two lines and P_c be the point in the datasets DS_1 and DS_2 that is closest to P_i . Let x_1 be the distance of P_i from P_c and x_2 is the distance of P_i from P_{\max} , then:

$$x_1 = |P_i - P_c| \quad (5)$$

$$x_2 = |P_i - P_{\max}| \quad (6)$$

Total distance, x , is given as:

$$x = x_1 + x_2 \quad (7)$$

Finally, the new focused pixel, P , is calculated as:

$$P = w_1 P_c + w_2 P_{\max} \quad (8)$$

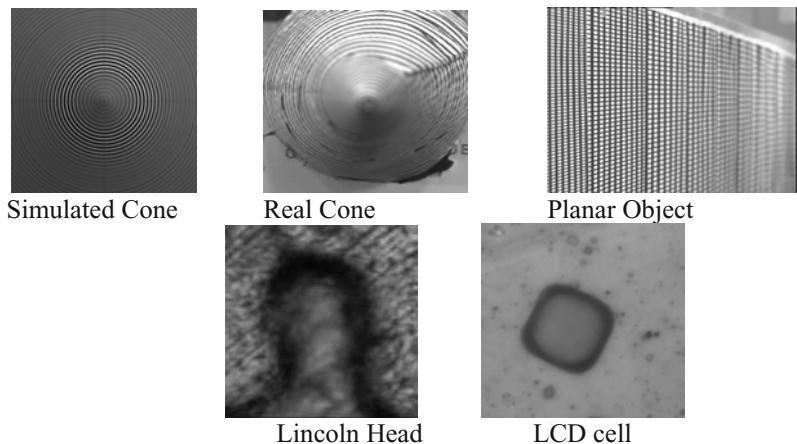
$$\text{where } w_1 = \frac{x_2}{x} \text{ and } w_2 = \frac{x_1}{x}$$

Hence more weight is given for the point closest to the intersection point as compared to P_{\max} . In case P_{\max} is closest point then the second closest point is taken as P_c .

4 Results and Discussion

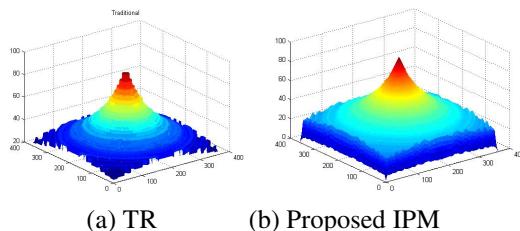
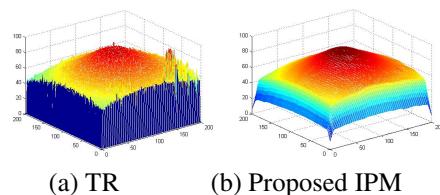
The proposed algorithm is applied on a sequence of 97 simulated cone images, 97 real cone images, 87 real planar object images, 68 microscopic coin (Lincoln head on US penny) images and 60 microscopic TFT-LCD cell images. The resolution of the images is 360x360 pixels for cone images, 200x200 for planar object and 300x300 pixels for the microscopic images. Fig 1 shows one of the frames for all objects.

The real cone was taken from the CCD camera system. The real cone object was made of hardboard with black and white stripes drawn on the surface so that a dense texture of ring patterns is viewed in the images. The absolute displacement between two consecutive image frames was about 0.03mm. Similarly, a sequence of 97 simulated cone images were generated corresponding to 97 lens positions. The third object used was a slanted planar object. A total of 87 images were acquired corresponding to 87 lens positions for this object, which has only horizontal slope and equally spaced strips to make a poor texture.

**Fig. 1.** Test Objects

Apart from the normal images acquired by CCD camera, we also acquired microscopic object images using microscopic image acquisition setup. Using the motor driver to control the step motor, object is displaced in z-direction. Using this experimental setup, we obtained images for 2 objects, namely, Lincoln head part in one cent coin and TFT-LCD cell.

Figures 2 to 6 show the depth maps of simulated cone, real cone, slanted planar object, microscopic coin head and microscopic TFT-LCD cell respectively. The difference

**Fig. 2.** Depth maps for the simulated cone**Fig. 3.** Depth maps for the real cone

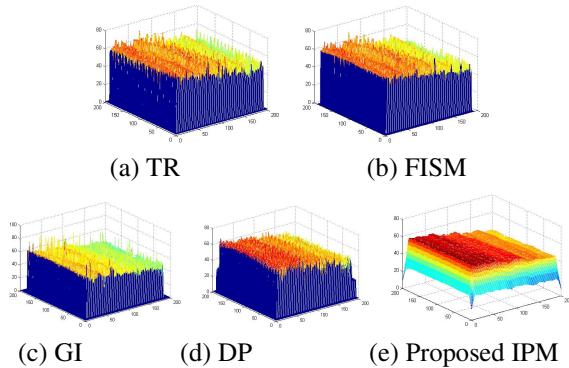


Fig. 4. Slanted planar object

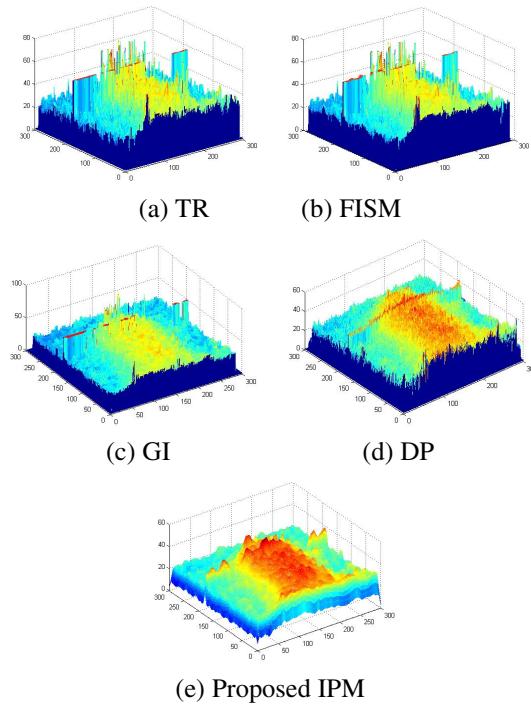


Fig. 5. Microscopic coin object

in the figures of the proposed method with other methods is quite clear. We can see that the surface of depth map from the proposed Intersection Point Method (IPM) is much smoother as compared to that of the other methods, namely, the Traditional

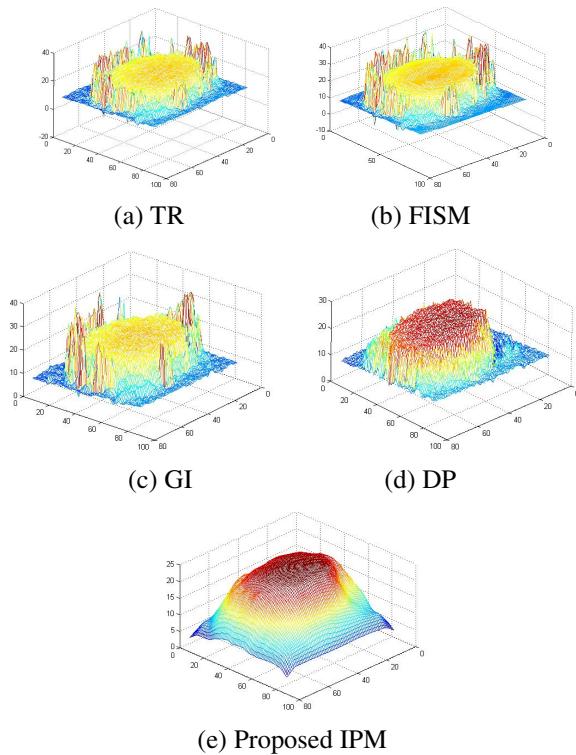


Fig. 6. Microscopic TFT-LCD Cell

(TR) method, Focused Image Surface Method (FISM), Gaussian Interpolation (GI) method and Dynamic Programming (DP) method. For details of these methods, please see section 2.

5 Conclusions

In this paper, we presented a method for depth estimation using image focus based on the linear regression model. The intersection point is found for two lines that are approximated using least squares linear regression and the new value for the depth map is calculated by assigning weights to points based on their distance from the intersection point. As for computational efficiency, this method is computationally more efficient than FISM however less efficient than GI and DP.

Acknowledgement

This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korean government (MOST) (No. R01-2007-000-20227-0).

References

1. Helmli, F.S., Scherer, S.: Adaptive Shape from Focus with an Error Estimation in Light Microscopy. In: 2nd Int'l Symposium on Image and Signal Processing and Analysis, Croatia, pp. 188–193 (2001)
2. Nayar, S.K., Nakagawa, Y.: Shape from focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16(8), 824–831 (1994)
3. Malik, A.S., Choi, T.-S.: A Novel Algorithm for Estimation of Depth Map using Image Focus for 3D Shape Recovery in the Presence of Noise. *Pattern Recognition* 41(7), 2200–2225 (2008)
4. Subbarao, M., Choi, T.-S.: Accurate recovery of three dimensional shape from image focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17(3), 266–274 (1995)
5. Malik, A.S., Choi, T.-S.: Application of Passive Techniques for Three Dimensional Cameras. *IEEE Transactions on Consumer Electronics* 53(2), 258–264 (2007)
6. Choi, T.S., Asif, M., Yun, J.: Three-dimensional shape recovery from focused image surface. In: IEEE International Conference of Acoustics, Speech and Signal Processing, vol. 6, pp. 3269–3272 (1999)
7. Asif, M., Choi, T.-S.: Shape from focus using multilayer feedforward neural network. *IEEE Transactions on Image Processing* 10(11), 1670–1675 (2001)
8. Ahmad, M.B., Choi, T.-S.: A Heuristic approach for finding best focused shape. *IEEE Transactions on Circuits and Systems for Video Technology* 15(4), 566–574 (2005)
9. Malik, A.S., Choi, T.-S.: Consideration of Illumination Effects and Optimization of Window Size for Accurate Calculation of Depth Map for 3D Shape Recovery. *Pattern Recognition* 40(1), 154–170 (2007)

Improved Phase Correlation Matching

Javed Ahmed and M. Noman Jafri

Image Processing Centre, NUST Military College of Signals, Rawalpindi, Pakistan
`{javed, mnjafri}@mcs.edu.pk`

Abstract. Phase correlation based template matching is an efficient tool for translation estimation which is in turn required for the image registration and the object tracking applications. When a template of an object is phase correlated with the search image, the resulting correlation surface is supposed to contain a sharp peak corresponding to the location of the object in the search image. However, the resulting surface also contains various *false* peaks which are sometimes *higher* in magnitude than the *true* peak. In order to solve the problem, we present an efficient and effective preprocessing technique that extends the images with new pixels having decaying values. The technique is compared with two recent methods on cluttered, noisy, blurred, and slightly rotated scenes. The results show that the proposed method outperforms both of them, especially when the object is away from the central region in the image.

1 Introduction

Precise estimation of the amount of translation of one image with respect to another is an important stage in an image registration application. Similarly, accurate localization of an object in a video frame is a critical step in an object tracking application. Both of these problems can be considered as a matching problem, which can be efficiently solved by cross-correlating the two images. However, the standard correlation is illumination sensitive and produces un-normalized response with wide peaks resulting in inaccurate matching [1]. On the contrary, phase correlation (PC) is significantly robust to illumination variation and offers a normalized response with a sharp peak at the best match location [12345]. The PC between a search image, s , and a template, t , is computed as:

$$c = \mathfrak{F}^{-1} \left(\frac{\mathfrak{F}(s)}{|\mathfrak{F}(s)|} \cdot \frac{\mathfrak{F}(t)^*}{|\mathfrak{F}(t)|} \right) \quad (1)$$

where $\mathfrak{F}(\cdot)$ and $\mathfrak{F}^{-1}(\cdot)$ are the 2-D DFT (discrete Fourier transform) and inverse DFT functions, respectively, the $|\cdot|$ operator computes the magnitude of every complex number in its input matrix, the asterisk (*) is the complex-conjugate operation, and the division and the multiplication are performed element-by-element. The normalization of the DFTs of s and t is usually called whitening of the signals [1].

If the PC is performed directly on the *original* signals, the result is the circular correlation, since the DFT assumes that each of the two signals is repeated

after its final element. In order to obtain the true PC, the signals must be appropriately zero-padded before computing their DFTs [3][6]. However, the zero-padding operation makes the spurious frequencies due to the discontinuities at the boundaries of the signals more evident, resulting in false peaks in the correlation response [4]. The problem occurs when any one of these false peaks has higher magnitude than that of the true peak.

Stone et al. [4] proposed to modulate (i.e. multiply element-by-element) the original images with a Blackman window [7] before zero-padding operation. The Blackman window smoothly attenuates the value of the image pixel depending on its position relative to the image center. Therefore, the discontinuities at the boundaries of the image are eliminated and the PC between the resulting images becomes significantly robust [4]. However, there is a disadvantage of this technique that it distorts the actual content of the image. That is, if the object of interest (e.g. a vehicle) lies significantly away from the center in the search image as shown in Fig. 1(left), the object becomes almost invisible in the modulated image as illustrated in Fig. 1(right). Similarly, the modulated template is also distorted, as depicted in Fig. 1. As a result, the modulated template does not match with the actual object as good as it does with the clutter at the central region, and the resulting PC response contains multiple false peaks at the center having magnitude higher than that of the true peak, as shown in Fig. 2(left). Thus, the phase correlation between the Blackman windowed images is failed to localize the actual object [Fig. 2(right)].

Keller et al. [5] suggested performing post-processing instead of pre-processing. That is, they phase correlated the original images and used a projection operator to restrict the support of the resulting surface. The projection operator zeros the correlation result beyond a rectangular region at the center. This approach unreasonably assumes that the translation is not significant between the two images being registered, or that the object of interest cannot lie away from the rectangular region at the center. Figure 3(left) shows the surface obtained after applying the projection operator on the PC performed between the original search image and template [Fig. 1(left)]. Since the object and the corresponding true peak were away from the rectangular support at the center, the true peak was set to zero and there remained only the false peaks around the center in the



Fig. 1. Blackman window modulation distorts images. (*left*) Original search image and template; (*right*) Modulated search image and template.

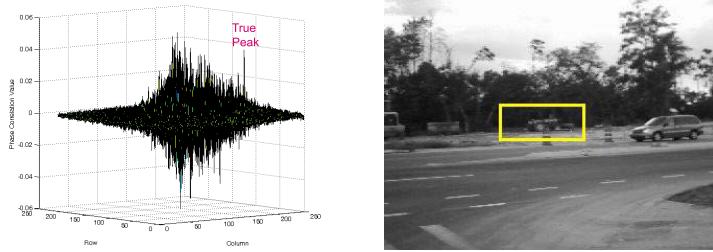


Fig. 2. Blackman window causes various false peaks in the correlation surface (if the object is away from the central region) resulting in false detection. (*left*) PC of the Blackman window modulated images indicating that the true peak is smaller than the false peaks and the highest false peak is at (105, 117) with magnitude only 0.0524; (*right*) Best-match rectangle with top-left position at (105, 117).

surface. Figure 3 (*right*) shows that the highest peak in the processed surface does not correspond to the vehicle.

In order to attenuate the false peaks and amplify the true peak, we propose an effective method that (a) slightly extends the search image and the template with new pixels having gradually decaying values (without changing the actual content of the images), (b) pads the extended images with zeros (without introducing any discontinuity), and (c) phase correlate the resulting images.

The next section describes the proposed technique of extending an image. Sect. 3 compares the proposed method with the two techniques mentioned in this section for cluttered, noisy, blurred, and slightly rotated scenes having reflection effect, and finally Sect. 4 concludes the paper.

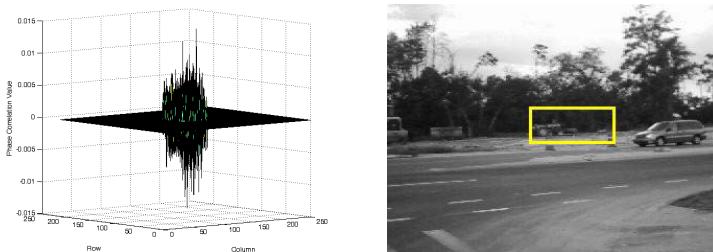


Fig. 3. The projection operator zeros the PC response [of the original images shown in Fig. 2(*left*)] beyond a rectangular region (typically 21×21) at the center, nullifying the true peak at the boundary, and resulting in a false detection. (*left*) The resulting surface shows the highest (false) peak at location (104, 141) with magnitude only 0.014; (*right*) Incorrect detection corresponding to the highest available peak.

2 Decaying Extension of an Image

We propose to extend an image using three steps as described in the following sub-sections.

2.1 Initializing the Extended Image

Let the original image be denoted by I having width, W , and height, H . We want to extend the image from every direction by δ pixels, as shown in Fig. 4(left). Thus, the size of the extended image, I_e , becomes $W_e \times H_e$, where $W_e = W + 2\delta$ and $H_e = H + 2\delta$. If $\delta < 4$, the discontinuities are not eliminated adequately, resulting in unsatisfactory attenuation of the false peaks. On the contrary, if $\delta \gg 4$, the discontinuities are eliminated very smoothly, but the computation time taken by the PC between the larger images is increased accordingly. In order to remain in the safe side, we set $\delta = 5$ in all of our experiments to have robust image matching without significantly increasing the computation time. We initialize every pixel in the extended regions with the value of its nearest boundary pixel in the original image. The process of initialization of the extended image can be described mathematically, as:

$$I_e(x, y) = \begin{cases} I(0, 0) & \text{if } 0 \leq x \leq \delta - 1 \text{ and } 0 \leq y \leq \delta - 1, \\ I(x - \delta, y - \delta) & \text{if } \delta \leq x \leq W + \delta - 1 \text{ and } \delta \leq y \leq H + \delta - 1, \\ I(0, H - 1) & \text{if } 0 \leq x \leq \delta - 1 \text{ and } H \leq y \leq H + \delta - 1, \\ I(W - 1, 0) & \text{if } W \leq x \leq W + \delta - 1 \text{ and } 0 \leq y \leq \delta - 1, \\ I(W - 1, H - 1) & \text{if } W + \delta \leq x \leq W_e - 1 \text{ and } H + \delta \leq y \leq H_e - 1. \end{cases} \quad (2)$$

2.2 Generating the Decaying Weights

Once the extended image is initialized, we proceed to generate the decaying weights for the new pixels. We obtain the weights using a Gaussian function because of its smooth and symmetric behavior.

Consider a \mathbb{R}^K column vector, where $K = 2\delta + 1$, containing the Gaussian weights: $g(k) = \exp[-(1/2)\{(k - \delta)^2/\sigma^2\}]$, where $k = 0, 1, \dots, K - 1$, and the standard deviation, σ , controls the spread of the Gaussian function. If the value of σ is too large, the function will go down too slow, and the boundary elements will be too high from zero, resulting in a sharp discontinuity during the zero-padding operation. On the contrary, if its value is too low, the function will go down abruptly to near-zero value, resulting in a discontinuity even before the zero-padding operation. We want to have the function, that goes down smoothly and becomes near-zero right at the boundary elements of the vector. We achieve this objective by computing an appropriate value of σ automatically according to the size of the vector, using: $\sigma = 0.3[(K/2) - 1] + 0.8$, as in [8]. Then, we split the vector g into its top and bottom halves (i.e. \mathbb{R}^δ vectors), as: $g_t(k) = g(k)$ and $g_b(k) = g(k + \delta + 1)$, $\forall k = 0, 1, \dots, \delta - 1$.

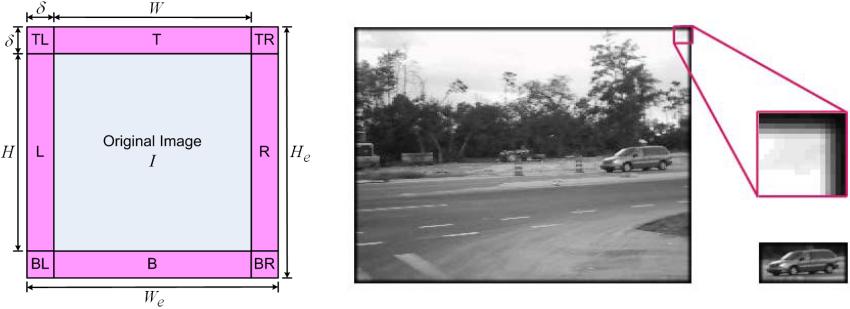


Fig. 4. Decaying extension of an image. (*left*) Structure of an extended image, I_e , where T = Top, L = Left, R = Right, and B = Bottom. (*right*) Search image and template from Fig. II(*left*) extended using $\delta = 5$ with new pixels having decaying magnitude down to near-zero, as can be observed easily in the zoomed-in corner region.

Now, consider a $\mathbb{R}^{K \times K}$ matrix containing the weights obtained by a 2D Gaussian function, as:

$$g(k_x, k_y) = \exp \left[-\frac{1}{2} \left\{ \frac{(k_x - \delta)^2 + (k_y - \delta)^2}{\sigma^2} \right\} \right], \quad \forall k_x, k_y = 0, 1, \dots, \delta - 1. \quad (3)$$

The value of σ is the same as computed previously. Then, we split the 2-D Gaussian function into its four quarters (i.e. $\mathbb{R}^{\delta \times \delta}$ matrices). The top-left, top-right, bottom-left, and bottom-right quarters are obtained as: $G_{tl}(k_x, k_y) = G(k_x, k_y)$, $G_{tr}(k_x, k_y) = G(k_x + \delta + 1, k_y + \delta + 1)$, $G_{bl}(k_x, k_y) = G(k_x, k_y + \delta + 1)$, and $G_{br}(k_x, k_y) = G(k_x + \delta + 1, k_y)$, $\forall k_x, k_y = 0, 1, \dots, \delta - 1$.

2.3 Applying the Decaying Weights

We multiply the g_t vector element-by-element with every column in the **T** patch in the initialized extended image [see Fig. II(*left*)], g_b vector with every column in the **B** patch, g_t^T row-vector with every row in the **L** patch, g_b^T row-vector with every row in the **R** patch, G_{tl} matrix with the **TL** patch, G_{tr} matrix with the **TR** patch, G_{bl} matrix with the **BL** patch, and G_{br} matrix with the **BR** patch. Figure II(*right*) illustrates the extended versions of the search image and template shown in Fig. II(*left*). It may be noted that all the content of the original images is intact in the extended images, and that only the pixels at the external regions are gradually getting more and more black as depicted by the zoomed-in top-right corner of the extended search image.

3 Experimental Results

When we zero-padded the extended versions of the search image and the template [Fig. II(*right*)] and phase correlated the resulting images with each other,

we obtained a single dominant peak in the correlation surface [Fig. 5(left)] corresponding to the true location of the vehicle in the search image [Fig. 5(right)]. We can see that the proposed technique has drastically attenuated all the false peaks in the PC response. Moreover, since the absolute of the sum of all the peaks in the PC response is always 1 [4], the energy released by the attenuating false peaks combines with that of the true peak, resulting in significant amplification of the true peak. This is a significant improvement introduced by the proposed approach in the PC response as compared to the response obtained using Blackman window (Fig. 2) or projection operator (Fig. 3).

Furthermore, we compared the proposed approach with the two techniques also for the template matching based object tracking application on numerous real videos. The size of the rectangular support of the projection operator [5] was enlarged from 21×21 to one-half of the frame size in order to give it some more advantage. We show two image sequences due to space constraint. Figure 6 illustrates a very noisy, shaky, and blurred video of a ground vehicle recorded from a UAV (unmanned aerial vehicle). The ground vehicle is being continuously translated and slightly rotated, and there is a noticeable reflection effect at the end of the video. The template – shown at the bottom-left of every frame – is kept constant to create a more challenging scenario. We can observe that only the proposed approach is successful in tracking the ground vehicle persistently in this video even when the vehicle is far from the central region of the frame. Figure 7 depicts the results of the three techniques on a cluttered and noisy video (OneShopOneWait2cor.avi) from CAVIAR database¹. Since the PC is only for estimating the translation between two images, it is not supposed to handle the significant variation in the appearance of the walking woman to be tracked in this video. Therefore, we smoothly update the template in each iteration of the techniques being compared, using α -tracker template updating method [9] (with $\alpha = 0.25$).

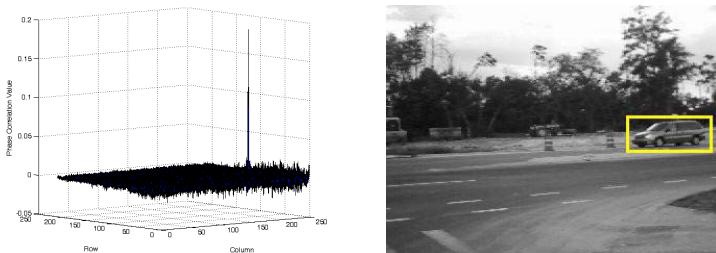


Fig. 5. Effect of the decaying extension of the images. (*left*) Phase correlation of the extended versions of the search image and the template that shows a single dominant peak at the true location of the best match, (111, 237), having magnitude amplified to 0.1836, and all the false peaks attenuated to a very small magnitude; (*right*) Correct detection of the vehicle at (111, 237) position.

¹ Available at <http://homepages.inf.ed.ac.uk/rbf/CAVIAR/>

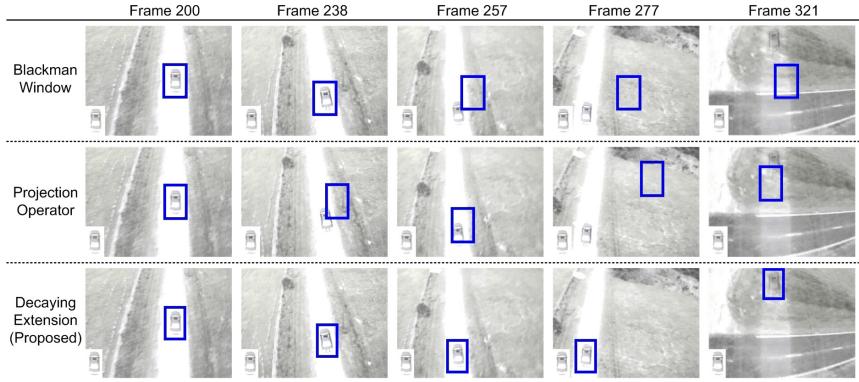


Fig. 6. Tracking a ground vehicle in a noisy, shaky, and blurred UAV (unmanned aerial vehicle) video containing slight rotations and significant reflection effect, when the template (shown at the bottom left of every frame) is not updated. Only the proposed method tracks the ground vehicle persistently.

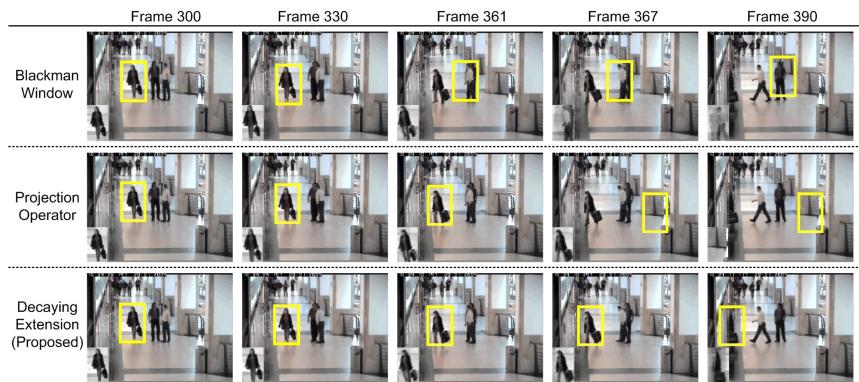


Fig. 7. Tracking a woman in a noisy and cluttered video from CAVIAR database, when the template (shown at the bottom-left of every frame) is iteratively updated. The proposed method tracks the woman robustly, while the other methods do not.

The adaptive template is shown at the bottom-left of every frame. It can be again observed that the proposed method outperforms the other two techniques in tracking the walking woman robustly even when she goes away from the central region of the frame.

We also compared the computation time of each technique by using a fixed 512×512 search image and varying the template size from 25×25 to 250×250 . In all cases, the proposed method took about as much time as the standard PC or the projection operator based PC, and only one-half of the time taken by the Blackman window based PC.

4 Conclusion

Phase correlation of the original images sometimes results in false peaks having magnitude higher than that of the true peak. This is mainly due to the discontinuities resulting from the mismatch between the elements at the opposite boundaries of each image and the zero-padding operation (necessary for obtaining the true correlation). We propose an efficient and effective technique that extends the original images with new pixels having decaying values, so that the discontinuities are eliminated without distorting the actual scene. Our method outperforms two previous techniques, even if the object is slightly rotated and away from the central region of the image suffering from significant reflection, noise, blur, and clutter. The accurate and robust pixel-level registration result from the proposed approach can definitely increase the true positive rate and the precision of the sub-pixel registration technique (described in [10]), that increase the number of samples in the phase correlation response and fill the in-between samples using some interpolation method.

References

1. Foroosh, H., Zerubia, J., Berthod, M.: Extension of phase correlation to sub-pixel registration. *IEEE Trans. on Image Processing* 11(3), 188–200 (2002)
2. Kuglin, D., Hines, D.: The phase correlation image alignment method. In: Proc. IEEE Conference on Cybernetics and Society, pp. 163–165 (1975)
3. Manduchi, R., Mian, G.: Accuracy analysis for correlation-based image registration algorithms. In: Proc. ISCAS 1993: Int'l Symposium on Circuits and Systems, vol. 1, pp. 834–837 (1993)
4. Stone, H., Tao, B., McGuire, M.: Analysis of image registration noise due to rotationally dependent aliasing. *Journal of Visual Communication and Image Representation* 14(3), 114–135 (2003)
5. Keller, Y., Averbuch, A., Miller, O.: Robust phase correlation. In: Proc. 17th IEEE International Conference on Pattern Recognition (2004)
6. Ahmed, J., Jafri, M., Shah, M., Akbar, M.: Real-time edge-enhanced dynamic correlation and predictive open-loop car-following control for robust tracking. *Machine Vis. and App.* 19(1), 1–25 (2008)
7. Oppenheim, A., Schafer, R., Buck, J.: *Discrete-Time Signal Proc.*, 2nd edn. Prentice Hall, Englewood Cliffs (1999)
8. Bradski, G., Kaehler, A., Pisarevsky, V.: Learning-based computer vision with open source computer vision library. *Intel Tech. J.* 9(3), 118–131 (2005)
9. Blackman, S., Popoli, R.: *Design and Analysis of Modern Tracking Systems*, 2nd edn. Artech House, Boston (1999)
10. Tian, Q., Huhns, M.N.: Algorithms for subpixel registration. *Comp. Vision, Graph., and Image Proc.* 35, 220–223 (1986)

A New Approach to Estimate Fractal Dimension of Texture Images

André R. Backes and Odemir M. Bruno

Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)
Avenida do Trabalhador São-carlense, 400
13560-970 São Carlos SP Brazil
backes@icmc.usp.br,
bruno@icmc.usp.br

Abstract. One of the most important visual attributes for image analysis and pattern recognition is the texture. Its analysis allows to describe and identify different regions in the image through pixel organization, performing a better image description and classification. This paper presents a novel approach for texture analysis, based on calculation of the fractal dimension of binary images generated from a texture, using different threshold values. The proposed approach performs a complexity analysis as the threshold values changes, producing a texture signature which is able to characterize efficiently different texture classes. The paper illustrates the novel method performance on an experiment using Brodatz images.

Keywords: Fractal Dimension, Texture Analysis, Complexity.

1 Introduction

Texture is one of the most important visual attribute used for pattern recognition. It is characterized by repetition, exactly or with small variations, of a gray-scale or colors pattern over a region, and this pattern is straight connected with physical properties of the surface of the object [123].

Its analysis allows to remark regions with same reflectance characteristics, and thus, the same colors in a determined combination of bands. This makes texture an excellent region descriptor, contributing for a more accurate process of image recognition, description and classification [134].

Although its great importance in several applications, its analysis process demands a high level of sophistication and computer complexity. Moreover, it is a term which depends on our perception and still does not have a precise definition [14].

Fractal dimension [567] is a measure of how complex an object is. In texture, this complexity is characterized by pixels organization, i.e., the complexity is straight connected to the visual aspect of the texture. So, fractal dimension allows to quantify a texture in terms of homogeneity, and makes possible a comparison between different textures [78].

In this work we propose a novel approach for texture characterization based on complexity analysis. Opposite to compute the complexity straight from the gray-scale texture, our main interest is to obtain a texture signature based on the complexity of binary images computed from the original texture using different thresholds. The complexity analysis is performed through application of the Box-Counting fractal dimension [9,10,6,7] over each binary image. The analysis of the obtained signature is performed using Linear Discriminant Analysis (LDA), which furnishes a reliable classification [11,12].

The paper is divided as follows. A brief review of complexity and fractal dimension is presented in Section 2. In Section 3, we present a novel approach to apply the Box-Counting fractal dimension over gray-scale images in order to compute a texture signature. In Section 4 we set up an experiment to compare the performance of the proposed texture signature to 400 images of 40 different Brodatzs texture classes. Section 5 shows the results yielded by the proposed method, while in Section 6, the conclusions about the method are discussed.

2 Fractal Dimension

Complexity can be understood as a measure of how irregular an object is. Its is a term largely used on the literature, but, in spite of its importance, it lacks of an exact definition. An interesting way to estimate the complexity of an object is through fractal dimension.

Fractal dimension (*FD*) is a measure of how irregular or complex an object is. Different from topological dimension, an integer value which describes the number of dimensions of the space where the object is inserted, fractal dimension is a fractionary value which describes how much of the space an object occupies [6,7].

One of the most known and used methods to estimate the fractal dimension of an object is the Box-Counting method. Its due to characteristics like easy implementation and simplicity of the computation involved in the method. Its computation is performed overlaying an image $A \in R^2$ with a grid of squares of size r and counting the number of squares, $N(r)$, necessary to cover the object in the image (Figure 1) [9,6,7].

The estimative of Box-Counting fractal dimension is based on relation between the size r used in the grid of squares and the number of squares, $N(r)$, which contain a portion of the image object A :

$$FD = -\lim_{r \rightarrow 0} \frac{\log(N(r))}{\log(r)}.$$

From log-log plot of r versus $N(r)$, we achieve a curve with α slope where

$$FD = -\alpha$$

is the Box-Counting fractal dimension of the object A (Figure 2) [13,7,10].

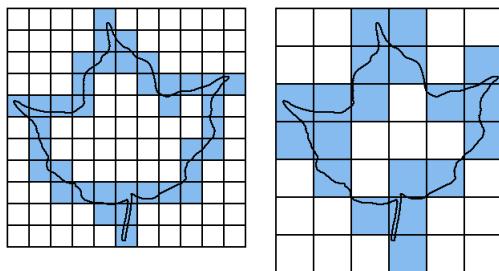


Fig. 1. Division of an image using the Box-Counting method for different r values

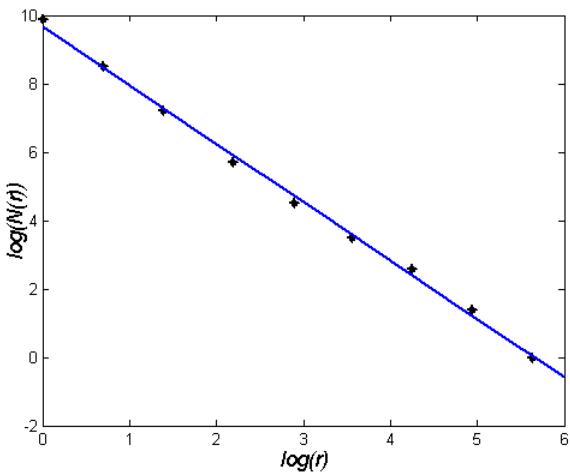


Fig. 2. Log-log plot of r versus $N(r)$

3 Proposed Method

In this work we propose a novel approach for texture characterization based on complexity analysis. The basic idea of the method is to compute the fractal dimension of binary images. These images are generated as we apply different thresholds in the texture under analysis. So, it's possible to compose a texture signature that describes how image complexity changes as we increase the threshold used.

Given a texture image $A \in R^2$ and a threshold value T_i , $T_i \in T$, a binary image A_{T_i} is generated applying an δ_{T_i} transformation as follow:

$$A_{T_i} = \delta_{T_i}(A) = \forall a \in A \begin{cases} a_{T_i} = 0, & \text{if } a < T_i \\ a_{T_i} = 1, & \text{if } a \geq T_i \end{cases}$$

The complexity analysis is performed by applying the Box-Counting fractal dimension over the image A_T :

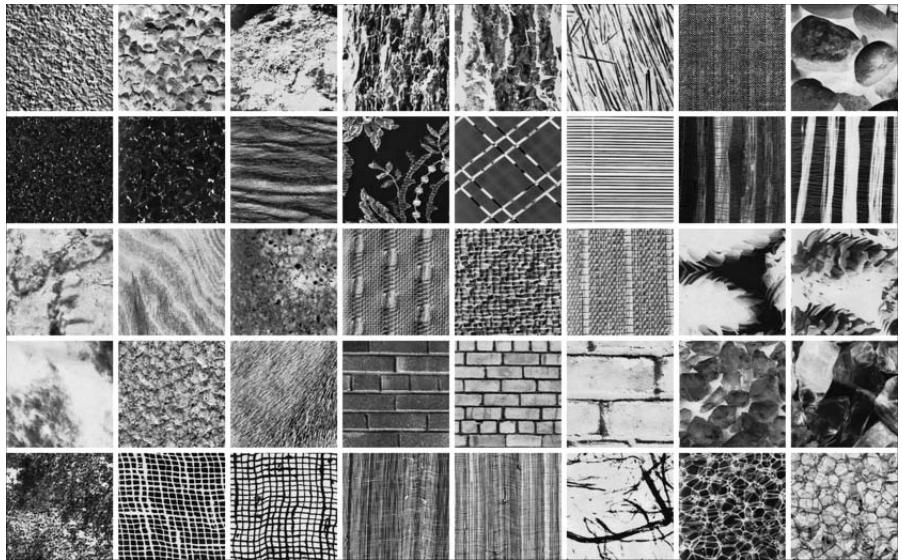


Fig. 3. One example of each of the 40 Brodatzs classes considered. Each image has 200 × 200 pixels and 256 gray levels.

$$FD_{T_i} = - \lim_{r \rightarrow 0} \frac{\log(N_{T_i}(r))}{\log(r)}.$$

Considering different threshold values T_i , its is possible to compute a texture signature $\psi(A)$:

$$\psi(A) = [FD_{T_1}, FD_{T_2}, \dots, FD_{T_N}] \quad i \in 1 \dots N,$$

where N is the number of thresholds considered for the texture characterization.

4 Experiments

In order to evaluate the proposed method an experiment has been carried out. The method has been calculated for different images from the book of Brodatz [4], which are largely used as benchmark for texture analysis in computer vision and image processing literature. For the experiment, 40 classes of images with 10 samples each has been employed. Each image sample is a portion of 200 × 200 pixels with 256 gray levels extracted from the original Brodatz image. One example of each of these 40 classes is showed in Figure 3.

The analysis step has been carried out applying over the texture signatures a Linear Discriminant Analysis (LDA), a well known method for estimating a linear subspace with good discriminative properties [1]. The idea of the method is to find a feature space where the data projection presents variance between the classes larger than the variance within the classes. LDA is a supervised method,

so it needs the class definitions for the estimation process [11][12]. A leave-one-out cross-validation scheme has been also employed.

An important issue in the method is the threshold set used to compute the texture signatures. The threshold set that better discriminate a texture set is that one which considers characteristics of the textures under analysis. So, a threshold selection approach has been used as follows:

1. For each texture in the database, compute its respective histogram;
2. Compute the mean histogram from the histograms of all textures. This mean histogram shows how the gray levels are distributed over the textures in the entire database;
3. Select N different threshold levels by applying the Otsu's method of multilevel threshold [15][16] over the mean histogram;

This approach allows us to select the thresholds values where the between-class variance of the different gray levels in the database is maximum.

5 Results

To achieve a threshold set that better discriminate a texture is a difficult task. The choice of the thresholds values, as also the quantity of threshold values, depends on the characteristics presents in the textures under analysis. Multilevel Otsu's method is a reliable way to choose the optimal thresholds of an texture image by maximizing the between-class variance with an exhaustive search [15][16]. When applied over the mean histogram of the texture database, multilevel Otsu's method furnishes a thresholds set related to the gray-levels distribution of the entire database, which allows to compose a more efficient texture signature $\psi(A)$.

Figure 4 shows the success rate of the proposed method as a function of the number of thresholds N computed by multilevel Otsu's method. As expected, we note a increase in the success rate as we increase the number of thresholds levels used to composed the texture signature $\psi(A)$. This is due to the additional information about the texture complexity that is added to the signature vector. Each threshold value represents a gray-level where the homogeneity of the texture changes. Complexity analysis methods, such Fractal Dimension, measures the homogeneity of an image through its complexity. Multilevel Otsu's method combined with Fractal Dimension method allows to create a signature vector which describes the most relevant changes in texture complexity. In fact, using 70 threshold levels, the proposed method was able to classify correctly 97.00% of the texture samples.

Figure 5 shows the texture complexity as a function of the threshold for four given texture samples of two different texture classes. Although fractal dimension value decreases as we increase the applied threshold, it is important to note that it occurs at different rates for each texture class. It evidences the use of the proposed feature vector $\psi(A)$ as a feasible texture signature.

For comparison intentions, the proposed method have been compared with traditional texture analysis methods found in the literature. These descriptors are: Fourier descriptors [17], Co-ocurrence matrices [18] and Gabor filters

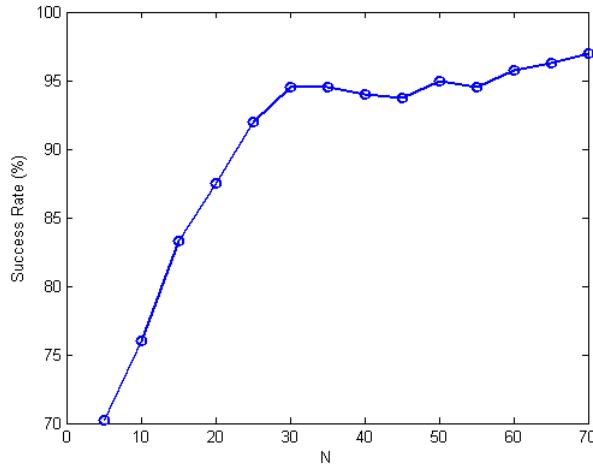


Fig. 4. Success rate as a function of the number of threshold levels (N) considered. Best classification (97.00%) is achieved when using $N = 70$.

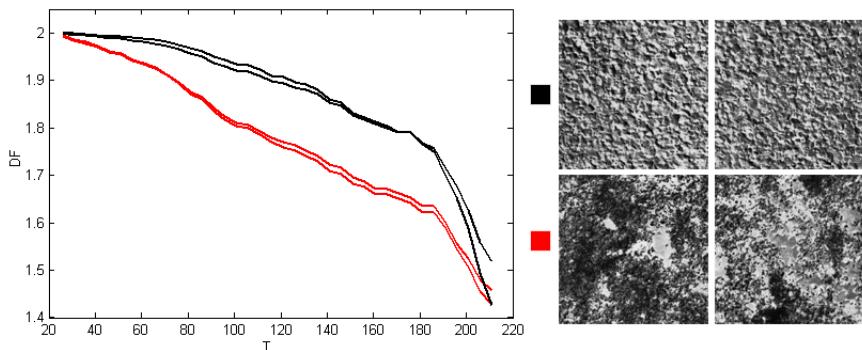


Fig. 5. Example of signatures computed for two different texture classes

[19][20][21]. Several different versions have been found on literature for these methods, however this paper considers just their conventional implementation.

Fourier descriptors: in the experiments, Fourier descriptors for texture analysis are composed of a feature vector sufficing the energy of the 99 most meaningful coefficients of the Fourier Transform. As low frequencies are placed at the center of the spectrum (after performing a *shifting* operation), each one of the coefficients corresponds to the sum of the spectrum absolute values placed to a radial distance from the center of the bi-dimensional transformation.

Co-ocurrence matrices: basically, they are joint probability distributions between pairs of pixels at a determined distance and direction. In this paper

Table 1. Comparison results for different texture methods

Method	Images correctly classified	Success rate (%)
Co-ocurrence matrices	330	82.50
Fourier	351	87.75
Gabor Filters	381	95.25
Proposed Method	388	97.00

approach, distances of 1 and 2 pixels with angles of -45° , 0° , 45° , 90° were used. Energy and entropy were computed from resulting matrices, totalizing a set of 16 descriptors. A non-symmetric version has been adopted in experiments.

Gabor filters: The 2-D Gabor filter is, basically, a bi-dimensional gaussian function moduled with an oriented sinusoid in a determined frequency and direction. This procedure consists of convolving an input image by a family of Gabor filters, which presents several scales and orientations of the same original configuration. Among numerous tests performed, the best results were achieved using a family of 16 filters (4 rotation filter and 4 scale filters), with frequency lower than 0.01 and superior than 0.3. Definition of the individual parameters of each filter follows mathematical model presented in [22].

Table I shows the yielded results presented by each method. We realize that, though the proposed method uses a simple texture discrimination approach, the yielded results overcome traditional texture methods. Moreover, the proposed method also presents the lowest error rate for class, what indicates a great capacity of discriminating different texture classes.

6 Conclusion

We have presented a new approach of texture feature extraction based on the Box-Counting fractal dimension. The method uses several threshold values applied over a texture image to compute the changes in its complexity. Thresholds are computed using the Otsu's method for multilevel threshold, which allows to choose the optimal thresholds by maximizing the between-class variance. The proposed approach outputs a curve strictly related to the configuration of the texture and, consequently, to its complexity. An experiment using the proposed signature and linear discriminant analysis to classify a set of Brodazt texture have been performed. The results show the great potential of the method to be used as a texture analysis methodology.

Acknowledgements

A.R.B. acknowledges support from FAPESP (2006/54367-9). O.M.B. acknowledges support from CNPq (303746/2004-1).

References

1. Ebert, D., Musgrave, K., Peachey, D., Perlin, K., Worley.: *Texturing and Modeling: A Procedural Approach*. Academic Press, London (1994)
2. Wang, L., Liu, J.: Texture classification using multiresolution markov random field models. *Pattern Recognition Letters* 20(2), 171–182 (1999)
3. Unser, M.: Texture classification and segmentation using wavelet frames. *IEEE Trans. Image Processing* 4(11), 1549–1560 (1995)
4. Emerson, C.W., Lam, N.N., Quattrochi, D.A.: Multi-scale fractal analysis of image texture and patterns. *Photogrammetric Engineering and Remote Sensing* 65(1), 51–62 (1999)
5. Chaudhuri, B.B., Sarkar, N.: Texture segmentation using fractal dimension. *IEEE Trans. Pattern Anal. Mach. Intell* 17(1), 72–77 (1995)
6. Schroeder, M.: *Fractals, Chaos, Power Laws: Minutes From an Infinite Paradise*. W.H. Freeman, New York (1996)
7. Tricot, C.: *Curves and Fractal Dimension*. Springer, Heidelberg (1995)
8. Lange, G.D., Marks, W.B.: Fractal methods and results in cellular morphology - dimensions, lacunarity and multifractals. *Journal of Neuroscience Methods* 69(2), 123–136 (1996)
9. Coelho, R.C., Costa, L.F.: The box-counting fractal. dimension: Does it provide an accurate subsidy for experimental shape characterization? if so, how to use it? In: *Anais do Sibgrapi*, vol. 95, pp. 183–191 (1995)
10. Li, J., Sun, C., Du, Q.: A new box-counting method for estimation of image fractal dimension. In: *International Conference on Image Processing*, pp. 3029–3032 (2006)
11. Everitt, B.S., Dunn, G.: *Applied Multivariate Analysis*, 2nd edn. Arnold (2001)
12. Fukunaga, K.: *Introduction to Statistical Pattern Recognition*, 2nd edn. Academic Press, London (1990)
13. Falconer, K.J.: *Fractal geometry: mathematical foundations and applications*, 288 p. Wiley, Chichester (1990); CALL NUM: QA614.86 .F35 1990
14. Brodatz, P.: *A Photographic Album for Arts and Design*, vol. 1. Dover Publishing Co., Toronto (1966)
15. Liao, P.S., Chen, T.S., Chung, P.C.: A fast algorithm for multilevel thresholding. *J. Inf. Sci. Eng* 17(5), 713–727 (2001)
16. Otsu, N.: A threshold selection method from gray level histograms. *IEEE Trans. Systems, Man and Cybernetics* 9, 62–66 (1979); minimize intra and inter class variance
17. Azencott, R., Wang, J.P., Younes, L.: Texture classification using windowed fourier filters. *IEEE Trans. Pattern Anal. Mach. Intell.* 19(2), 148–153 (1997)
18. Haralick, R.M.: Statistical and structural approaches to texture. *Proc. IEEE* 67(5), 786–804 (1979)
19. Jain, A.K., Farrokhnia, F.: Unsupervised texture segmentation using Gabor filters. *Pattern Recognition* 24(12), 1167–1186 (1991)
20. Daugman, J., Downing, C.: Gabor wavelets for statistical pattern recognition. In: Arbib, M.A. (ed.) *The Handbook of Brain Theory and Neural Networks*, pp. 414–419. MIT Press, Cambridge (1995)
21. Idrissa, M., Achteroy, M.: Texture classification using gabor filters. *Pattern Recognition Letters* 23(9), 1095–1102 (2002)
22. Manjunath, B.S., Ma, W.Y.: Texture features for browsing and retrieval of image data. *IEEE Trans. Pattern Anal. Mach. Intell* 18(8), 837–842 (1996)

Digital Signal Processing Techniques for Gene Finding in Eukaryotes

Mahmood Akhtar^{1,2}, Eliathamby Ambikairajah², and Julien Epps²

¹Centre for Health Informatics, University of New South Wales,
45 Beach St, Coogee NSW 2034, Australia

²School of EE&T, University of New South Wales,
Sydney 2052, Australia

mahmood@unsw.edu.au, ambi@ee.unsw.edu.au, j.epps@unsw.edu.au

Abstract. In this paper, we investigate the effects of window shape and length on a DFT-based method for gene and exon prediction in eukaryotes. We then propose a new gene finding method which combines the selected time-domain and frequency-domain methods, by employing the most effective DNA symbolic-to-numeric representation examined to date in conjunction with suitable window shape and length parameters and a signal boosting technique. It is shown herein that the new method outperforms major existing approaches. By comparison with the existing methods, the proposed method reveals relative improvements of 15.1% to 55.9% over different methods in terms of prediction accuracy of exonic nucleotides at a 5% false positive rate using the GENSCAN test set.

Keywords: DNA, periodicity, discrete Fourier transforms, signal boosting.

1 Introduction

It is well-known that deoxyribonucleic acid (DNA) is the material of heredity in most living organisms, and consists of genic and intergenic regions. Eukaryotes differ from prokaryotes in that their genes are further divided into relatively small protein coding segments known as *exons*, interrupted by non-coding spacers known as *introns*. Eukaryotic gene finding is a significant open problem in the field of DNA sequence analysis. The problem is difficult mainly due to the noncontiguous and non-continuous nature of genes. Furthermore, often the intergenic and intronic regions make up most of the genome. For example, in human genome the exonic fraction is as low as 2%.

The conversion of DNA nucleotide symbols (i.e., A, C, G, and T) into discrete numerical values enables novel and useful DSP-based applications for the solution of different sequence analysis related problems such as gene finding [1]. In recent years, a number of schemes have been introduced to map DNA nucleotides into numeric values [2]. The binary or Voss representation [3] is a popular scheme, which maps the nucleotides A, C, G, and T into the four binary indicator sequences $x_A[n]$, $x_C[n]$, $x_G[n]$, and $x_T[n]$ showing the presence (e.g. 1) or absence (e.g. 0) of the respective nucleotides. For example, for a DNA sequence $x[n] = \text{AGTTCTACCGAGC...}$, the binary

indicator sequences for each base type would resemble: $x_A[n] = \{1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, \dots\}$, $x_C[n] = \{0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, \dots\}$, $x_G[n] = \{0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, \dots\}$, $x_T[n] = \{0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, \dots\}$, where $x_A[n] + x_C[n] + x_G[n] + x_T[n] = 1$, and n represents the base index. In exons, the occurrence of identical nucleotides in identical codon (e.g. triplet encoding protein) positions is the basis for a periodicity of three interpretation in these regions [4]. The period-3 behaviour of exons has been widely used to identify these regions using DSP-based methods such as the discrete Fourier transform (DFT) [1], [5]–[7], time-domain algorithms [8], etc.

The discrete Fourier transform (DFT), the most commonly used method for spectrum analysis of a finite-length numerical sequence $x[n]$ of length N , is defined as:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi nk}{N}}, \quad 0 \leq k \leq N-1 \quad (1)$$

Equation (1) can be used to calculate DFTs for four binary indicator sequences (i.e., $X_A[k]$, $X_C[k]$, $X_G[k]$, and $X_T[k]$). The periodicity of 3 in exon regions of a DNA sequence suggests that the DFT coefficient corresponding to $k = N/3$ (where N is chosen to be a multiple of 3) in each DFT sequence should be large [1]. Various DFT based spectral measures exploiting the period-3 behaviour of exons for the identification of these regions have been proposed. The spectral content (SC) measure [5] combines the individual DFTs (i.e., $X_A[k]$, $X_C[k]$, $X_G[k]$, and $X_T[k]$) to obtain a total Fourier magnitude spectrum of the DNA sequence, as follows:

$$SC[k] = \sum_m |X_m[k]|^2, \quad m \in \{A, C, G, T\} \quad (2)$$

The spectral rotation (SR) measure [6] rotates four DFT vectors $X_A[k]$, $X_C[k]$, $X_G[k]$ and $X_T[k]$ clockwise, each by an angle equivalent to the average phase angle value in coding regions, to make all of them ‘point’ in the same direction. The SR measure also divides each term by the corresponding phase angle deviations to give more weight to exonic distributions. The feature

$$SR[k] = \left| \sum_m \frac{e^{-j\mu_m}}{\sigma_m} X_m[k] \right|^2, \quad m = \{A, C, G, T\} \quad (3)$$

has been used for the detection of exons. The SR measure has been shown [6] to give better performance than the SC (2) measure at a 10% false positive gene detection rate. The paired and weighted spectral rotation (PWSR) measure [7] incorporates a statistical property of eukaryotic sequences, according to which introns are rich in nucleotides ‘A’ and ‘T’ whereas exons are rich in nucleotides ‘C’ and ‘G’. This information leads to an alternative to the well-known period-3 behavior of exons. In this method, the DNA sequences are first converted into two binary indicators (i.e., $x_{A,T}[n]$ and $x_{C,G}[n]$). Using training data from DNA sequences of the same organism, the means μ_m and standard deviations σ_m of the distributions of DFT phase angle averaged over coding regions (i.e., one phase angle value for one coding region) are calculated. Weights w_m based on the frequency of occurrence of nucleotides ‘A or T’ and ‘C or G’ in coding regions of the training data are also calculated. The expression given in (4) can then be used as a feature, along one direction of the DNA sequence:

$$PWSR_i[k] = \left| \frac{e^{-j\mu_{A-T}} \cdot w_{A-T} \cdot X_{A-T}[k] + e^{-j\mu_{C-G}} \cdot w_{C-G} \cdot X_{C-G}[k]}{\sigma_{A-T}} \right|^2 \quad (4)$$

where l = forward (F) and reverse (R) directions of DNA sequence, and $X_m[k]$ ($m = A-T, C-G$) are the sliding DFT windows of two indicator sequences. The expression in (4) has been used in the reverse direction of the same DNA sequence (i.e., due to paired indicators, DFT in reverse direction of the same DNA strand is equivalent to DFT on its complementary strand). The PWSR measure is the sum of forward and reverse measures:

$$PWSR[k] = PWSR_F[k] + PWSR_R[k] \quad (5)$$

The PWSR measure has been shown [7] to give better performance than the SC (2) and SR (3) measures at 10%, 20%, and 30% false positive nucleotide detection rates.

Time-domain algorithms [8] use prefiltering of the four binary indicator sequences (i.e., pass them through a second order resonant filter with centre frequency of $2\pi/3$) to remove spectral components at $2\pi k/3$, $k \in \mathbb{I}$, $k \neq 1$, which arises from the application of correlation-based approaches to a binary indicator sequence at a base-domain lag of 3. The resultant non-binary numerical sequences, emphasizing the period-3 behavior of DNA sequences but de-emphasizing the other components, are input to the average magnitude difference function (AMDF) or time domain periodogram (TDP) algorithms. The AMDF for a discrete signal $x[n]$ as a function of the period k , is defined as:

$$AMDF[k] = \frac{1}{N} \sum_{n=1}^N |x[n] - x[n-k]| \quad (6)$$

where N is the window length. Practically, the AMDF will produce a deep null if significant correlation exists at period $k = 3$.

Despite the existence of these approaches and also data-driven approaches, the accuracy of gene prediction still needs to be improved. We address this shortcoming herein, investigating the effects of window shape and length parameters and proposing a new DSP-based gene finding method. The remainder of this paper is organized as follows. In Section 2, effects of window shape and length on the DFT-based gene and exon prediction are investigated. A new digital signal processing-based gene finding method is then proposed in Section 3. Finally, the proposed and existing methods are compared in Section 4.

2 Effects of Window Shape and Length

Most existing signal processing-based gene prediction methods use sliding window attributes to maintain a reasonable base-domain resolution. The DFT-based SC measure [5], SR measure [6], are well-known examples. All these authors use a conventional rectangular shaped window of length of 351 bp, with the argument that the data window should be reasonably long or few hundred base pairs long [5]-[6]. Datta and Asif [9] claim an improved DFT based period-3 detection using the Bartlett data window compared with the rectangular window of same size (i.e., $N = 351$). How does window shape and/or window size affects the performance of exon prediction methods? What could be an optimal window size for a given set of sequences? Herein, we

address these questions through an investigation of the suitability of different window shapes and length for period-3 exon detection.

2.1 Database and Evaluation Metrics

The DNA sequences were first converted into four binary indicator sequences, using the Voss representation [3], as discussed in Section 1. The DFT-based SC measure (2) was then used for the period-3 detection, using different window shapes (with fixed length of 351) such as rectangular, Bartlett, Hanning, Hamming, Blackman, Kaiser and Gaussian (e.g. see Fig. 1). In each case, the entire HMR195 dataset [10] containing 195 mammalian gene sequences, was used. These sequences have maximum length 200,000 bp, and the ratio of human:mouse:rat sequences is 103:82:10, with the mean exon length of 208 bp. The receiver operating characteristic (ROC) curve measure was used to compare the performance of SC measure using different window shapes. An ROC curve explores the effects on true positive (*TP*) and false positive (*FP*) as the position of an arbitrary decision threshold is varied, and plots the *TP* as a function of *FP* of exonic and intronic nucleotide separation method for varying decision threshold values, where *TP* is the number of coding nucleotides correctly predicted as coding and *FP* is the number of non-coding nucleotides predicted as coding. One way of characterizing this result as a single number is to calculate the area under the ROC curve (AUC), with larger areas indicating more accurate detection. Empirically, we found shape parameter values $\alpha = 1.5$ and $\beta = 2$ respectively for the Gaussian and Kaiser windows, more suitable for the DFT-based gene prediction. The Kaiser window was further investigated for different lengths.

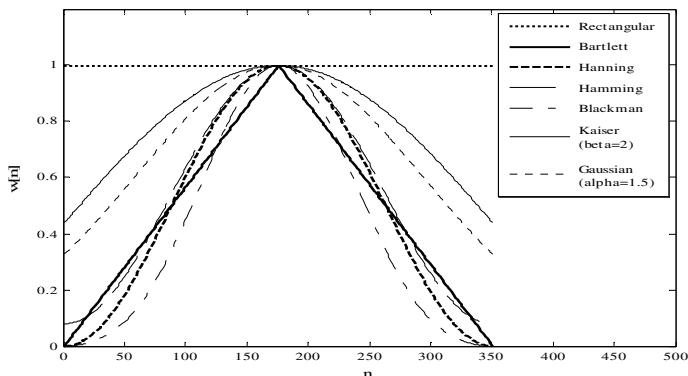


Fig. 1. Time-domain representation of different window shapes

2.2 Results and Discussion

Table 1 summarizes the AUC results for different window shapes (for a constant window length of 351). It is clear that the window type does affect the performance of the DFT based period-3 detection method. Results show that the Kaiser and Gaussian windows are better than other types, whereas the Blackman is poorer. The performance of rectangular, Bartlett, and Hamming windows seems comparable. It is very

interesting to see from the time-domain representation of different window types (in Fig. 1) that the extremely suppressed edges in different shapes (e.g. Blackman, and Hanning) result in a loss of useful DNA sequence information, and are presumably the main reason for poor gene prediction results using these types. On the other hand, abrupt truncation of sequences as in rectangular type, causes leakage of the DFT power into adjacent frequencies. This work suggests that a window shape with moderately suppressed edges (e.g. Gaussian $\alpha = 1.5$, and Kaiser $\beta = 2$ in our case) would be the more suitable choice for the DFT-based gene prediction.

Table 1. AUC results using different window types

Window Type	Area under ROC curve (AUC)	Window Type	Area under ROC curve (AUC)	Window Type	Area under ROC curve (AUC)
Rectangular	0.8008	Hamming	0.7979	Kaiser ($\beta = 2$)	0.8059
Bartlett	0.7985	Blackman	0.7816	Gaussian ($\alpha = 1.5$)	0.8056
Hanning	0.7923				

The AUC as function of beta (β) for different length Kaiser windows is shown in Fig. 2. Regardless of the window length, the optimum performance of the Kaiser window lies around a value of $\beta = 2$. A similar experiment (results not shown) suggests that the Gaussian window give the optimum performance for a value of $\alpha = 1.5$ (approximate). Clearly, the effect of window length on DFT-based exon prediction is stronger than that due to the window shape parameter. It can be observed that due to an average exon length of 208 bp in HMR195 dataset, the Kaiser window with a length of 201 points performs better than those with lengths 351, 501, and 651, acknowledging results in [11], using a different window type on a completely different

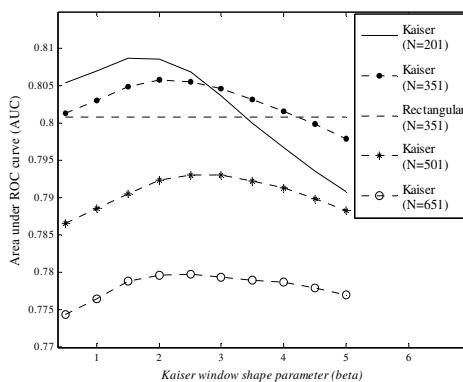


Fig. 2. AUC vs. β plot for different lengths of Kaiser window, using the HMR195 dataset

dataset. This too suggests that an unnecessary larger window size not only requires longer computation time, but it also gives poor performance for different window types, compromising the base-domain resolution.

3 Proposed Method for Gene Prediction

The paired numeric method has recently been shown [2] to be the most effective DNA symbolic-to-numeric mapping scheme for the DFT-based gene prediction. Results and discussion presented in Section 3, suggest that a Kaiser window ($\beta = 2$) of length approximately equal to the average exon length of the given dataset is the more suitable choice of window shape and length parameters. Furthermore, the signal boosting technique has been shown [12] a successful post-processing of output signals to enhance genomic protein coding regions and suppress the non-coding regions. Herein, we first modify the frequency-domain PWSR measure (5) by employing the best DNA representation, a suitable window shape and length, and the signal boosting technique. The modified PWSR measure is then combined with the time-domain AMDF method (6) to improve gene prediction accuracy of existing methods. The block diagram of the proposed optimized time-frequency hybrid setup is shown in Fig. 3.

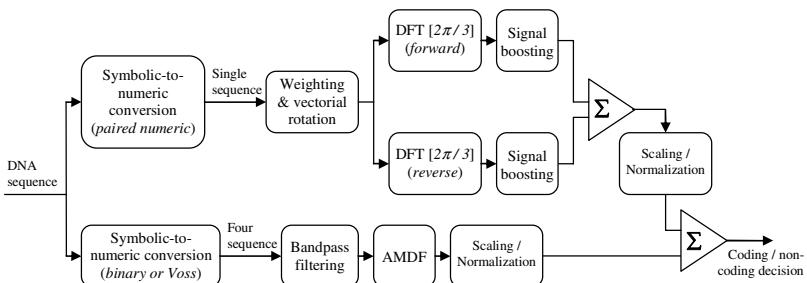


Fig. 3. Block diagram for the proposed optimized time-frequency hybrid method

Before applying the PWSR method, the DNA sequence is first converted into numeric values using the paired numeric representation [2], previously shown empirically to be the best available mapping scheme for the gene and exon prediction problem [2]. Weights w_m based on the frequency of occurrence of bases ‘A or T’ and ‘C or G’ in coding regions of the training data are then assigned. A reduction in DFT processing is achieved after symbolic-numeric conversion by applying the spectral rotation and weighting of the PWSR measure before rather than after the DFT processing, recognizing that the DFT is a linear transform. The shape and length of the DFT window is another important performance parameter, investigated in Section 2, and a Kaiser window (with $\beta = 2$) length of 150 base pairs is used herein, assuming human genomic sequences. The recently proposed signal

boosting technique [12] is then applied to the forward and reverse DFT features, to enhance their values in protein coding and suppress them in non-coding regions. According to the signal boosting technique, the protein coding regions are treated as the ‘signal’, while non-coding regions are treated as the ‘noise’, and a gain factor $\Gamma(m)$ is calculated as the ratio of a short-term average signal energy $P(m)$ to the estimate of the noise floor level $Q(m)$ for $m = 1, 2, \dots, M$, where M is the length of DNA sequence. The boosted signal $\hat{X}(m)$ is then calculated as [12]:

$$\hat{X}(m) = \Gamma(m) X(m) = \frac{P(m)}{Q(m)} X(m) \quad (7)$$

where $X(m)$ is the period-3 detection DFT feature. Finally, the forward and reverse signal boosted PWSR features are combined with an unweighted sum. The resultant features are then used as an optimized PWSR feature for discrimination of coding and non-coding nucleotides. Finally, we combine ‘time’ and ‘frequency’ domain methods, similar to [7]. A simple fusion approach is employed, in which the features from each method are normalized to the range $[0, 1]$ and combined with an unweighted sum.

4 Evaluation

4.1 Database and Evaluation Metrics

Two datasets consisting of human genomic sequences (e.g. GENSCAN learning and test sets [13]) were employed for the training and testing of different methods. A constant window size of 351 was used for the existing DFT-based SC, SR, and PWSR measures, as suggested in their original descriptions [5]-[7]. A frame size of 117 was used for the AMDF method, similar to [8]. In implementations of the SR, PWSR and the proposed method, prior information (frequency of nucleotide occurrence weights and angular mean and deviation values) was obtained from the GENSCAN learning set. The discrimination power of all methods was measured and compared at the nucleotide level, using evaluation measures such as ROC curves, AUC, and percentage of exonic nucleotides detected as false positives, similar to [7].

4.2 Gene Prediction Results

From the ROC curve and area under ROC curve results summarized in Fig. 4 and Table 2, we see that the proposed method outperforms the existing time-domain, frequency-domain, and combined time-frequency measures, giving consistently improved exonic nucleotide detection and the largest area under ROC curve. The proposed method reveals relative improvements of 55.9%, 49.3%, and 26.4% respectively over the SC, SR, and PWSR measures in the detection of exonic nucleotides at a 5% false positive rate. Furthermore, the proposed method gives relative improvements of 27.3% and 15.1% respectively over the AMDF and existing time-frequency hybrid (TFH) measure [7] in the detection of exonic nucleotides at a 5% false positive rate. Although the improvements over existing methods at a 20% or

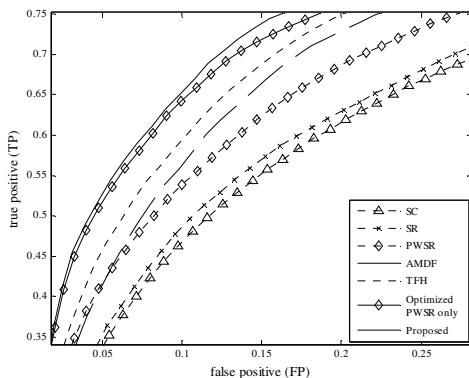


Fig. 4. ROC plot using GENSCAN test set

Table 2. Summary of results using GENSCAN test set

Method	Area under ROC curve	% of exonic nucleotides detected as false positive				
		5%	10%	15%	20%	30%
SC	0.7778	33.8	46.7	55.2	61.6	71.0
SR	0.7800	35.3	48.6	57.0	62.9	72.4
PWSR	0.8123	41.7	53.8	62.5	68.7	77.3
AMDF	0.8338	41.4	56.2	66.5	72.9	81.7
TFH	0.8448	45.8	59.5	68.8	74.9	81.6
Optimized PWSR only	0.8532	51.6	64.3	71.8	76.3	82.6
Proposed	0.8650	52.7	65.4	73.8	78.3	84.8

larger false positive rate are more modest, results at low false positive rates are more significant, due to the high likelihood of false positives resulting from the low exonic fraction in eukaryotic genomes.

5 Conclusion

We have investigated the effects of window shape and length parameters on DFT-based gene and exon prediction. This revealed that the effect of window length is stronger than that of window shape, and the optimum window length for the given dataset is approximately equal to the average length of exons. We have also proposed a new gene prediction method which employs the most effective DNA representation examined to date in conjunction with suitable window shape and length parameters and a signal boosting technique. Using the GENSCAN test set of human gene sequences; the proposed method outperforms all existing methods in this comparison. Future work may combine this optimized signal processing method with data-driven methods to advance the state of the art in detection of exonic/intronic end-point signals (e.g. acceptor/donor splice sites, start/stop codons).

References

1. Anastassiou, D.: Genomic Signal Processing. *IEEE Signal Proc. Mag.* 18(4), 8–20 (2001)
2. Akhtar, M., Epps, J., Ambikairajah, E.: On DNA Numerical Representations for Period-3 Based Exon Prediction. In: 5th IEEE Workshop on Genomic Signal Processing and Statistics, Tuusula, Finland (2007)
3. Voss, R.F.: Evaluation of Long-range Fractal Correlations and 1/f Noise in DNA Base Sequences. *Phy. Rev. Lett.* 68(25), 3805–3808 (1992)
4. Fickett, J.W.: Recognition of Protein Coding Regions in DNA Sequences. *Nucleic Acids Res.* 10, 5303–5318 (1982)
5. Tiwari, S., Ramaswamy, S., Bhattacharya, A., Bhattacharya, S., Ramaswamy, R.: Prediction of Probable genes by Fourier Analysis of Genomic Sequences. *Comput. Appl. Biosci.* 13, 263–270 (1997)
6. Kotlar, D., Lavner, Y.: Gene Prediction by Spectral Rotation Measure: A New Method for Identifying Protein Coding Regions. *Genome Res.* 18, 1930–1937 (2003)
7. Akhtar, M., Epps, J., Ambikairajah, E.: Time and Frequency Domain Methods for Gene and Exon Prediction in Eukaryotes. In: IEEE ICASSP, pp. 573–576 (2007)
8. Ambikairajah, E., Epps, J., Akhtar, M.: Gene and Exon Prediction using Time-Domain Algorithms. In: 8th IEEE Int. Symp. on Sig. Proc. and its Appl., pp. 199–202 (2005)
9. Datta, S., Asif, A.: A Fast DFT Based Gene Prediction Algorithm for Identification of Protein Coding Regions. In: IEEE ICASSP, pp. 653–656 (2005)
10. Rogic, S., Mackworth, A.K., Ouellette, B.F.: Evaluation of Gene-Finding Programs on Mammalian Sequences. *Genome Res.* 11(5), 817–832 (2001)
11. Akhtar, M., Ambikairajah, E., Epps, J.: Optimizing Period-3 Methods for Eukaryotic Gene Prediction. In: IEEE ICASSP, pp. 621–624 (2008)
12. Gunawan, T.S., Ambikairajah, E., Epps, J.: A Boosting Approach to Exon Detection in DNA Sequences. *IEEE Electronic Letters* 44(4), 323–324 (2008)
13. Burge, C.: Identification of Genes in Human Genomic DNA. PhD Thesis Stanford University, Stanford, CA, USA (1997)

A CAD System for Long-Bone Segmentation and Fracture Detection

Martin Donnelley, Greg Knowles, and Trevor Hearn

Flinders University, GPO Box 2100, Adelaide 5001, South Australia

{martin.donnelley, greg.knowles, trevor.hearn}@flinders.edu.au

Abstract. Medical imaging has advanced at a tremendous rate since x-rays were discovered in 1895. Today, x-ray machines produce extremely high-quality images for radiologists to interpret. However, the methods of interpretation have only recently begun to be augmented by advances in computer technology. Computer aided diagnosis (CAD) systems that guide healthcare professionals in making the correct diagnosis are slowly becoming more prevalent throughout the medical field. Detection of long-bone fractures is an important orthopaedic and radiologic problem, and it is proposed that a novel CAD system could help reduce the number of fractures missed during x-ray diagnosis. A number of image processing software algorithms useful for assisting the fracture detection process are described, and their accuracy evaluated on a database of fracture images from trauma patients. Incorporating these methods will further expand the capabilities of today's CAD systems, and result in more accurate diagnosis of fractures and a reduction of the fracture miss rate.

Keywords: Bone, CAD, Fractures, AMSS, Hough, Segmentation.

1 Introduction

Fractures of bone are a common affliction, accounting for approximately 20% occupancy of orthopaedic wards at any given time [1]. In the Australian population the number of fractures associated with age-related bone loss is increasing rapidly, with the number of non-hip fractures estimated to increase by 9% every 5 years until 2036 [2].

Trained radiologists generally identify abnormal pathologies including fractures with a relatively high level of accuracy. However studies examining reader accuracy [3] have shown that in some cases the miss rate can be as high as 30%, or even higher when reading x-rays containing multiple abnormalities. Accurate diagnosis of fractures is vital to the effective management of patient injuries. As a result, detection of long-bone fractures is an important orthopaedic and radiologic problem, and it is proposed that a novel CAD system could help lower the miss rate. This paper examines the development of such a system, for the detection of long-bone fractures.

To produce a CAD system for long-bone fracture detection, two major goals were identified. The first was to create a semi-automatic long-bone segmentation

method to separate the diaphysis from the proximal and distal epiphyses. The second was to create a semi-automatic long-bone fracture detection method that could identify fractures within the diaphyseal segment.

2 The Fracture Detection Algorithms

Our CAD system consists of four stages that are described in the following sections. For development and evaluation a set 50 long-bone images obtained from the Flinders Medical Centre Emergency Department was split into 6 training images and 44 testing images. All training images contained at least one fracture, while 38 of the 44 test images contained at least one fracture.

2.1 Long-Bone Edge Detection

Since the structures within an x-ray image were not easily computed, we used a scale-space approach to detect edges. Depending on the scale chosen, local disturbing detail could be removed and the features at that scale could be extracted. To compute features within an image $I(x, y)$, $(x, y) \in R$, scale-space theory proposes that the image be replaced by a smoothed version $I(x, y, t)$, where the degree of image smoothing is determined by the scale parameter t . After smoothing, the points with high gradient intensity $|\nabla I(x, y, t)|$ contain the most significant global information [4].

To meet all invariance requirements—including properties like Euclidean (translation and rotation) invariance—all multi-scale image analysis techniques should have the form of a curvature motion [5]. We were interested in a special case called the affine morphological scale space (AMSS):

$$\frac{\partial I}{\partial t} = |\nabla I| (curv(I))^{\frac{1}{3}}. \quad (1)$$

This case was chosen because it diffused purely in the direction orthogonal to the gradient, plus it was also the only case that produced an additional invariance termed affine invariance. A curve is said to be affine invariant if the coordinate system in which it is represented can change without affecting the relative geometry of the curve. An example image and the magnitude of the gradient—calculated using the AMSS—are shown in Figure 11.

2.2 Parameter Approximation

Rather than attempting to accurately segment the bone from the background using traditional segmentation algorithms, only the information about the location and orientation of the bone was determined. To extract position information, prior knowledge about expected human long-bone morphology was utilised. Regardless of their anatomical location, long-bone diaphyses containing no major

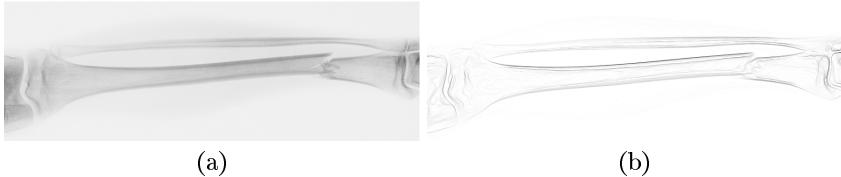


Fig. 1. (a) The original image, and (b) the magnitude of the gradient $|\nabla I(x, y, t)|$ calculated using the AMSS

pathology were typically very straight despite a large degree of natural variation in bone shape. This allowed the diaphyseal bone edges to be approximated by line straight parameters useful for both diaphysis segmentation and fracture detection.

Global information for approximating the long bones within the x-ray image was extracted from the magnitude of the gradient $|\nabla I(x, y, t)|$ using the normal parameterization of the Hough transform. Hence spatially extended patterns were converted into spatially compact features within the space of possible parameter values (ρ, θ) . Thus the problem was reduced from a difficult global feature analysis to a simpler local peak detection in the parameter space.

Peak detection in the resulting Hough accumulator H is a non-trivial exercise, especially with complicated input images. Simple peak detection algorithms that locate the absolute maximum or utilise a global threshold often work well for artificial images, but generally perform poorly when applied to real x-ray images. Even for an infinitesimally thin line, oversampling in the parameter space causes the peak in the accumulator to spread rather than remain localised to one cell. Thus, in practice the votes do not fall into uniform and compact regions, so the peaks are often poorly defined and spread in both the ρ and θ directions. Real world images containing thicker, non-straight lines—such as the edges $|\nabla I(x, y, t)|$ in a typical x-ray image—cause additional spread. The amount of expected spread was quantified using [6], and was used to iteratively locate the maximum elements in the accumulator and thus obtain a set of parameters (ρ_i, θ_i) to approximate the long-bone shaft.

Although the bone edges were approximated by the parameters (ρ_i, θ_i) , no information regarding the length or the endpoints of the lines was known. To accurately detect the long-bone shaft from the peak parameters, an endpoint detector based on [7] was created. In this technique, the path along each parameterised line was examined using a two stage thresholding technique. The first threshold was applied to the path so that only regions where the line matched the bone edge were retained, while the second was applied to ensure that only lines of a minimum length were retained. Figure 2b shows the lines detected by the long-bone shaft endpoint detector. The endpoints were located where the bone edge deviated sufficiently away from the line.

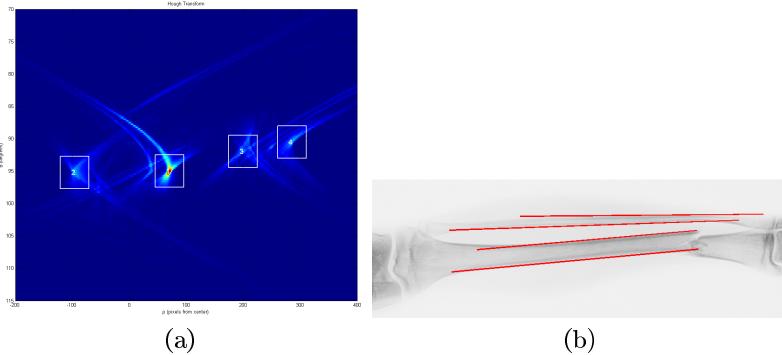


Fig. 2. (a) The Hough Transform of the image in Figure 1b, showing the detected peaks (and the estimate of the peak spread). (b) The lines detected using the long-bone shaft endpoint detector.

2.3 Diaphysis Segmentation

The first segmentation stage used the long-bone approximation parameters (ρ_i, θ_i) to calculate the best estimate of the bone centre-line. In general, centre-line detection was much more difficult for displaced fractures (those in which the centre-lines of the shaft segments on either side of the fracture remained colinear) than undisplaced fractures, and also for distal bones than proximal bones.

In cases where $i > 2$ it was necessary to determine which lines should be paired together. Lines that were to be paired should be located relatively close to each other, because the bone edges should in most cases be approximately parallel. This also meant that when correctly paired, the line endpoints should also be relatively close together. Thus it was proposed that the criterion for determining the correct pairings would be the minimum total distance between the line endpoints. So all the possible pairings were analysed and the combination that minimised this distance was chosen. When multiple centre-lines were detected for the distal long-bones, only the one along the axis of the wider bone was retained. An example of the relationship between the detected bone edges (red) and resulting bone centre-line (green) is shown in Figure 3a.

By definition the boundary between the diaphysis and the epiphysis is the point at which the long-bone shaft begins to widen and curve. Fortunately, the long-bone shaft endpoint detector was used to locate those points at which the bone deviated sufficiently away from the parameterised line fitted to the straight portion of the bone edge. Since the line endpoints on either side of the shaft often varied, it was appropriate to use an average measure, which was the same point at which the centre-line ended. The segmentation boundary at the diaphysis should also be orthogonal to the centre-line at that point. Therefore at both ends of every identified centre-line, an orthogonal line spanning the image was added, as shown dashed in blue in Figure 3a. Regardless of the number of centre-lines in the image, it was always the two orthogonal lines furthest apart that



Fig. 3. (a) Once the centre-lines (*green dashed*) had been located from all of the bone edges (*red solid*), lines orthogonal to the centre-line endpoints (*blue solid*) were added. From these lines the segmentation was performed. (b) The segmentation result.

corresponded to the diaphyses. An example diaphysis segmentation is shown in Figure 3b.

2.4 Fracture Detection Using Gradient Analysis

Long bone diaphyses consist almost exclusively of uniform cortical bone. X-rays of a typical diaphysis containing no abnormal pathology should exhibit gradient changes in the direction normal to the bone axis, due to the bone-tissue interface, and the variation in cortical wall thickness through which the x-rays must pass. The wall thickness and density change relatively slowly along the length of the long bone, so the gradient parallel to the bone edges is small. Large gradients occurring at angles not orthogonal to the bone edges are therefore indicative of some type of abnormality, and are detected by this algorithm.

In 2.2 line parameters to approximate the edges of the long bones were determined, and from this the magnitude and direction of the edges that should be present within the long-bone shaft are known. Removing these leaves behind only the abnormal gradients, such as those that belong to any fractures. To remove all normal regions from the image a composite measure of the magnitude $|\nabla I(x, y, t)|$ and direction $\phi(x, y, t)$ of the gradient of the smoothed image $I(x, y, t)$ at scale t , where t is suitably small for microscopic feature analysis is used.

The gradient composite measure (GCM) $C(x, y, t, \rho, \theta, p)$ is the product of the magnitude of the gradient and two scaling factors used to incorporate the direction of the gradient $\phi(x, y, t)$, with the long-bone shaft approximation parameters. The first scaling factor—called the importance rank $R(x, y, t, \theta, p)$ —measured how well the direction of the gradient matched the angle θ of the approximation lines, while the second—called the distance rank $D(x, y, \rho, \theta, p)$ —measured how close each pixel was to all the approximation lines:

$$C(x, y, t, \rho, \theta, p) = |\nabla I(x, y, t)| R(x, y, t, \theta, p) D(x, y, \rho, \theta, p). \quad (2)$$

The magnitude data was normalised to the range $[0, 1]$. However, the direction data was in the range $[0, 360^\circ]$ and had to be ranked in terms of which angle was most important, since large angles did not necessarily correspond to more

important regions. The importance rank $R(x, y, t, \alpha, p)$ was based on the direction of the gradient $\phi(x, y, t)$ at that pixel, a chosen transform angle α , and an importance weighting coefficient p . It was calculated such that the transform angles α and $\alpha \pm 180^\circ$ were assigned the minimum value 0, while orthogonal angles $\alpha \pm 90^\circ$ were assigned the maximum value 1. Fractures and other abnormalities were located by setting the transform angle α equal to the angle parameter θ_i of each of the approximating lines:

$$R(x, y, t, \theta, p) = \prod_{i=1}^e \left(1 - [\cos(\phi(x, y, t) - \theta_i)]^{2p} \right). \quad (3)$$

Similarly, the distance rank was a measure of how far a point in the image was from all of the long-bone shaft approximation lines. For a single line with parameters (ρ, θ) the distance rank for a point (x, y) was simply the length of the normal between that point and the approximation line. A power relationship p was again applied, so that regions close to the line were given a much higher importance than those further from the line. The distance rank was calculated for each point in the image using:

$$D(x, y, \rho, \theta, p) = \prod_{i=1}^e \left(|\rho_i - [(x - x_{origin}) \cos \theta_i + (y_{origin} - y) \sin \theta_i]|^{\frac{1}{p}} \right). \quad (4)$$

As shown in Figure 4a, the GCM did indeed indicate the location of the fracture, but some further processing was used to make the diagnosis clearer. Firstly a 9×9 median filter was used to remove salt and pepper type noise created by random matches between ϕ and θ in the GCM calculation. The regions most likely to correspond to a fracture were isolated by simultaneously applying two empirically chosen thresholds, only keeping the pixels that had a sufficiently large composite measure (greater than T_1), and also occurred in a large enough cluster to be significant (sum of the pixel intensities in the cluster was greater than T_2). Once these regions were identified they were enlarged to a user adjustable size using morphological dilation and their boundaries were marked on the original image. Figure 4b demonstrates the result.



Fig. 4. (a) The gradient composite measure (b) Clustering was applied to mark the identified regions

3 Evaluation

3.1 Long-Bone Edge Detection

Applying the algorithm to the testing images showed that the smoothing produced by the AMSS algorithm was contained within the boundaries of the bone, and that increasing the scale definitely resulted in greater smoothing.

3.2 Parameter Approximation

When the Hough Transform and peak detection methods were applied a total of 160 peaks were detected. Manual inspection of these peaks showed that 142 (88.8%) were in locations that correctly corresponded to the bone shaft edges. The incorrect peaks were dispersed across the 44 images such that only 12 images (27.3%) were completely correct (no false peaks). The highest rates of detection occurred in the humerus and femur, where all the detected peaks corresponded to points of interest on the shaft. Once all 185 peaks were correctly located, the line endpoint detector was applied. Manual inspection showed that only four of the 185 lines (2.2%) were incorrect. The incorrect lines were in four different images from the 44 image set, corresponding to 9.1% incorrect images. The incorrect lines all occurred as a result of good matches with edges belonging to other features within the image that happened to lie underneath the line of interest.

3.3 Diaphysis Segmentation

When tested, the centre-line detection algorithm correctly located 52 of the 56 centre-lines (93%) manually identified in the 44 images. In all four cases where a centre-line was not detected correctly, the cause was that the lines were not appropriately paired. In these cases the incorrect detection of the centre-line had little effect on the final segmentation, although this would not always be the case. Of the 71 segmentation boundaries in the test set, 59 were identified correctly, corresponding to a detection rate of 83%. In addition, in 33 out of the 44 images (75%), all the required segmentation boundaries were correctly detected, such that those images were segmented completely correctly. The most common cause of choosing the incorrect location for segmentation (6 cases) was the fracture disturbing the centre-line such that it ended prematurely, rather than continuing to the point where it would if the fracture were not present. Thus, segmentation was better in images where the bones were complete—that is, images containing subtle fractures with no displacement. In fact, when a subset of 24 images containing only subtle undisplaced fractures was used—rather than the complete set that also contained obvious grossly displaced fractures—the diaphysis segmentation rate was 100%.

3.4 Fracture Detection Using Gradient Analysis

Using a sensitivity analysis, values of $p = 5$, $T_1 = 0.008$ and $T_2 = 50$ were chosen. A manual comparison to the radiologist's results was performed, and the

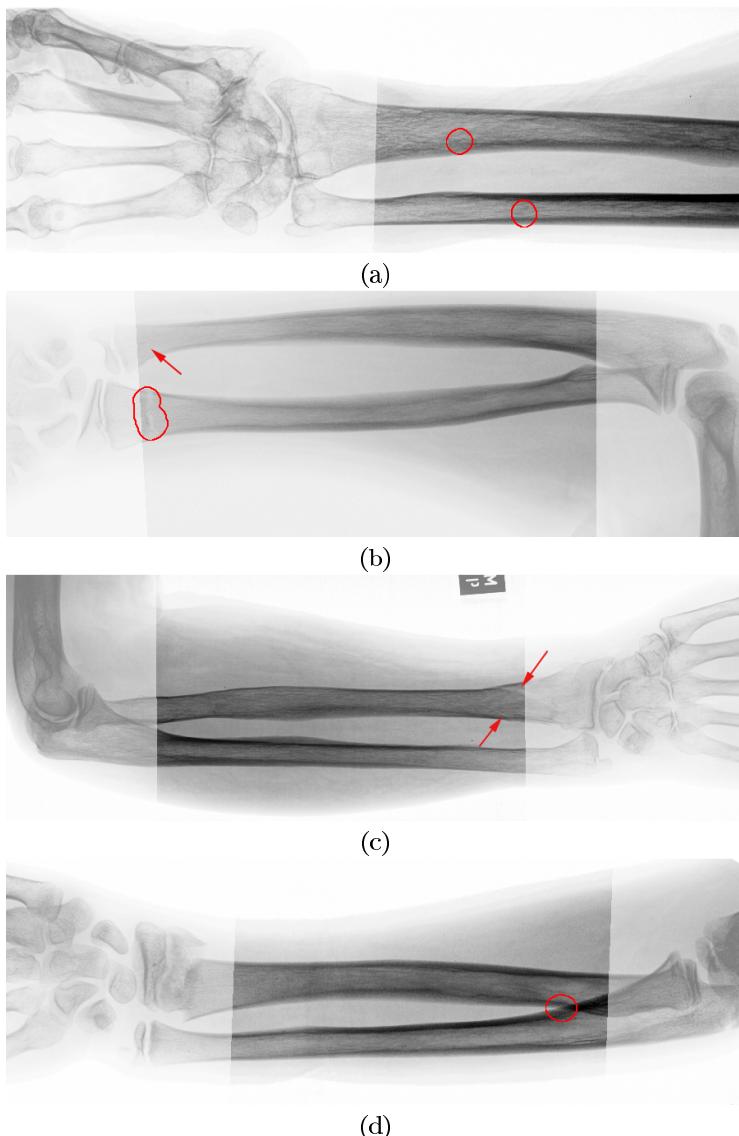


Fig. 5. Examples of the results produced by the fracture detection algorithm. (a) Two extremely subtle midshaft forearm fractures were detected correctly. (b) One subtle radius fracture was detected correctly, but the corresponding subtle ulna fracture (arrow) was missed. (c) Subtle radius and ulna fractures (arrows) were missed. (d) A false detection due to overlapping bones.

Table 1. (a) A confusion matrix showing the results produced by the algorithm on a sample of 47 fractures present in the 44 image test set. (b) The causes of false detection.

	Predicted Non-fracture	Predicted Fracture		
Non-fracture	1 (11%)	68		
Fracture	8 (17%)	39 (83%)		
(a)				
	Cause of false detection	Number	Individual %	Group %
Algorithm	Bone overlap	7	10.3%	27.9%
	Image artifact	12	17.6%	
Biological	Harris growth arrest line	5	7.3%	72.1%
	Feature not related to fracture	38	55.9%	
	Soft tissue injury	6	8.8%	
(b)				

number of true and false positive and negative results were recorded, along with the causes of any false positives. The test image set contained a total of 47 fractures, 39 of which were correctly detected by the algorithm, corresponding to a detection rate of 83% of all fractures. A summary of the results is shown in the confusion matrix in Table 1a. The causes of all the false positives were recorded, and were split into two categories: algorithm errors, and misinterpreted biological phenomena. Details are shown in Table 1b. Example results produced by the algorithm are shown in Figure 5.

4 Conclusion

Computer aided detection of midshaft long-bone fractures has not previously been examined. A method by which semi-automated long-bone shaft segmentation, along with fracture detection within the segmented region has been presented. In the test set, 83% of the diaphysis segmentation boundaries were correctly identified, and 83% of the fractures within those segmented regions were also detected correctly. These results will further expand the capabilities of today's CAD systems, and result in more accurate diagnosis of fractures and a reduction of the fracture miss rate.

References

1. Chipchase, L.S., McCaul, K., Hearn, T.C.: Hip fracture rates in South Australia: Into the next century. *Australian and New Zealand Journal of Surgery* 70(2), 117–119 (2000)
2. Sanders, K.M., Nicholson, G.C., Ugoni, A.M., Pasco, J.A., Seeman, E., Kotowicz, M.A.: Health burden of hip and other fractures in Australia beyond 2000. Projections based on the Geelong Osteoporosis Study. *Medical Journal of Australia* 170(10), 467–470 (1999)
3. Berlin, L.: Reporting the missed radiologic diagnosis: Medicolegal and ethical considerations. *Radiology* 192(1), 183–187 (1994)

4. Marr, D.: Vision: A computational investigation into the Human Representation and Processing of Visual Information. W. H. Freeman and Company, New York (1982)
5. Alvarez, L., Lions, P.L., Morel, J.M.: Image selective smoothing and edge detection by nonlinear diffusion. SIAM Journal of Numerical Analysis 29(3), 845–866 (1992)
6. Atiquzzaman, M.: Multiresolution Hough transform - An efficient method of detecting pattern in images. IEEE Transactions on Pattern Analysis and Machine Intelligence 14(11), 1090–1095 (1992)
7. Skingley, J., Rye, A.: The Hough transform applied to SAR images for thin line detection. Pattern Recognition Letters 6(1), 61–67 (1987)

Comparison of Image Restoration Methods for Bioluminescence Imaging

Smaïl Akkoul¹, Roger Ledee¹, Remy Leconge¹, Christophe Leger¹, Rachid Harba¹, Sabrina Pesnel², Stéphanie Lerondel², Alain Lepape², and Luis Vilcahuaman³

¹ Université d'Orléans, Institut PRISME, 12 Rue de Blois, BP 6744, 45067
ORLEANS CEDEX 2

² Centre d'Imagerie du Petit Animal, CIPA-CNRS, UPR n°44, 3 B, Rue de la Férollerie, 45071

³ Pontificia Universidad Católica del Perú, Facultad de Ciencias e Ingeniería, Lima, PERU
{Smail.Akkoul, Roger.Ledee, Remy.Leconge, Christophe.Leger,
Rachid.Harba}@univ-orleans.fr
{Sabrina.Pesnel, Lerondel, Lepape}@cnrs-orleans.fr

Abstract. Bioluminescence imaging is a recent modality to visualize biological effects, especially for small animals. However, the acquired images are degraded by diffusion and absorption phenomena from the tissue and by the acquisition system itself. In this paper, we use restoration methods to enhance the quality of bioluminescence images. We propose a model for image formation and an experimental determination of the PSF (Point Spread Function). Several restoration methods are compared on test images generated according to the model and on real data. This comparison is insured by using MSE (Mean Square Error) and two other quantitative criteria. Results showed that the statistical methods give more accurate restoration and are well adapted for Bioluminescence Imaging.

Keywords: Restoration methods, Bioluminescence Imaging, acquisition models.

1 Introduction

During the last years, research showed that there were numerous concordances between the human genome and that of primates and mice [1]. The advantage of working on mice is that they have a short generation time and convenient breeding conditions. It is possible to manipulate their human equivalent genes [2] to create models of human diseases [3]. But *in vivo* studies on these models need the development of novel investigation methods adapted to small animals.

Among the emergent techniques for *in-vivo* investigation of the small animal, bioluminescence imaging (BLI) [6, 7] proves to be very promising. Bioluminescence is a chemical reaction which produces visible light that can be observed in a variety of marine and terrestrial animals, more especially in the firefly [4, 5]. This imaging modality presents several advantages. It is simple, non-invasive and inexpensive. It is at the same time a functional and metabolic imaging of a gene expression [5]. Recently, it has been exploited in complement of other methods [8] to

monitor biological processes during time for tumors detection and drug therapy assessment and for follow-up of the neoplastic evolution [2, 3].

Nevertheless, diffusion and absorption of light by the tissues [6] make this method difficult to use in practice, mainly when quantitative imaging is required. Moreover, the acquisition process introduces a point spread function which makes that acquired image does not matching to the reality [8]. One of the possible solutions to correct these effects would be to use image restoration methods [10]. As bioluminescence imaging is quite a new technique, no such work has been carried out to our knowledge.

In this study, the image restoration process is used to improve the quality and recover the spatial resolution of bioluminescence images. We investigate several restoration methods and compare them on test images and real acquisitions.

This communication is organized as follow. In section 2, we present the material, the modeling of the system and all the methods of restoration studied as well as the parameters of comparison. The results obtained on synthesized images and real images are presented in section 3. Conclusions of this work form the last section.

2 Materiel and Methods

2.1 Acquisition System

Bioluminescent imaging offers many opportunities for non-invasive study using reporter genes. Tagged cells are labeled with reporter genes encoding luciferase enzymes in living small animal. To monitor the evolution of tagged cells, substrate luciferin is applied to the animal by injection. Those cells that express the luciferase transgene emit photons. The mice were anesthetized and placed onto warmed stage inside camera box. Our bioluminescent images have been acquired for one minute with a commercial unit (Hamamatsu, Macroscopic Imaging AEQUORIA), that comprises, a light-tight cabinet with a high sensitivity cooled CCD camera. The images obtained have a resolution of 512x512 pixels coded on 16 bits and 300 μm space resolution. An example of such images is presented in Fig. 1.

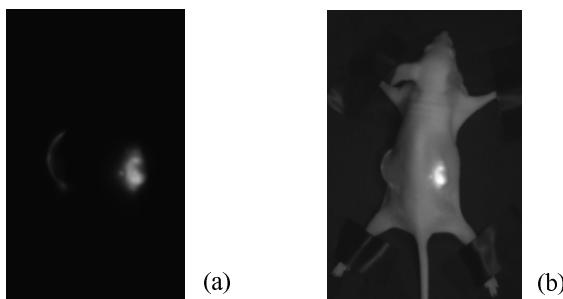


Fig. 1. (a) Example of a bioluminescence acquisition, (b) Fusion of the bioluminescence image and visible light image

2.2 Acquisition Model

The formation of the bioluminescence images results from the counting of photons emitted by a source radiating in all directions starting from the inside of the small animal and which is collected outside by the CCD camera. The light intensity of the sources is low and photon propagation is subject to both scattering and absorption phenomena.

In this work, we will only consider subcutaneous tumors which represent 80% of tumor locations in experimental cancerology. So, we can measure the PSF including instrumentation and skin effects. This is the first stage before introducing a model of diffusion and absorption of the skin in the restoration process. Bioluminescence imaging acquisition process is similar to that of nuclear medicine, astronomical and optical microscopic imaging. All of them use CCD camera and a low light phenomena is recorded. The quantum nature of light leads to a Poisson modeling of the signal emitted by the object. Recent works on these imaging modalities provide reviews of proposed models [20, 23, 25]. The camera is cooled (-75°C) to be extremely sensitive to the observed phenomena. Nevertheless, images are degraded by a noise due to CCD camera read-out, quantization, and dark current [9]. The additive noise is supposed to be a zero mean Gaussian noise of unknown variance σ^2 . We propose to adopt for the bioluminescence imaging a Poisson model. This model is expressed as:

$$i(x, y) = p([o(x, y) * h(x, y)]) + b(x, y), \quad (1)$$

where i : acquired image, o : real image, h : point spread function, p : Poisson process, $*$: convolution product and b : additive Gaussian noise.

2.3 Restoration Methods

The image restoration process is used to reduce the effects of the blur. This inverse problem has led to a large amount of work [10, 15, 16, 19, 21], the main difficulties are the determination of the PSF and the additive noise [22, 24]. Many classifications are proposed in the literature [23, 25]. In our work, we propose to classify used methods in four classes: linear, nonlinear, statistical, and blind methods.

The first class concerns the linear methods, like Wiener [10, 11] filter or that of Tikhonov-Miller [13, 23]. To decrease the sensitivity to noise of the inverse filter and its singularities, the Wiener filter minimizes the mean square error between the original image and the restored one. While it is easy to implement, it requires a good knowledge of the noise spectrum. It is expressed as:

$$MSE = E[(o(x, y) - \hat{o}(x, y))^2], \quad (2)$$

The Tikhonov-Miller filter found when minimizing the Tikhonov functional, which is the square difference between the acquired image and blurred estimate of the original object regularized by a Tikhonov energy bound. It is expressed as:

$$\phi(O(x, y)) = \arg \min \|i(x, y) - (h * o)(x, y)\|^2 + \lambda \|p * h\|^2, \quad (3)$$

where, the first term represents the fidelity to the data $I(x, y)$, and the second relates to the smoothing applied to the restored image: p is high pass filter, λ is the regularization parameter which represents the trade-off between the fidelity to the data and the smoothness of the restored image and the original image. This method is simple to implement, but it is very sensitive to errors in the PSF data used for estimation, leading to ringing artifact in the solution.

The second class includes the nonlinear methods. To solve the difficulties occurring within the first class, constraints were integrated in these linear methods. Historically, Jason-Van Cittert's algorithm is the first constraint method. It is an iterative algorithm that converges quickly. But it is not very effective in the case of images having a low *SNR*. Thereafter, this algorithm was improved by introduction of other constraints like limiting the intensity of the signal, for example. A nonlinear approach of Tikhonov-Miller algorithm was also proposed. The minimum of the Tikhonov function is reached iteratively by clipping the estimated negative intensities to zero after each iteration. Usually, a conjugate gradient (CG) method is used for the iterative search operation [13]. During iterations, the steepest descent in direction of minimum of the function is calculated. The solution is expressed as:

$$O_{k+1} = \begin{cases} O_k + s_k p_k & \text{if } O_k + s_k p_k \geq 0 \\ 0 & \text{otherwise} \end{cases}, \quad (4)$$

where s_k is the size of the optimal step and p_k is the conjugate gradient direction.

The improvement of this algorithm was the objective of many work but all are based on the assumption that the nature of the noise is Gaussian [13]. To take into account other physical models like Poisson process, statistical methods were proposed and are classified in the third class. Statistical methods are based on the Bayes's rule, which consist of constructing the conditional probability density relationship:

$$P(O/I) = P(I/O)P(O)/P(I), \quad (5)$$

where $P(O/I)$ denotes the probability density function (*pdf*) of O given I , $P(O)$ denotes the *pdf* of the real image, and $P(I/O)$ denotes the a priori probability. The objective is to find an estimate of the original image which maximizes the posterior probability. The maximization of the right part of the equation gives a solution O . This is obtained by only the maximization of the density $P(I/O)$ over O (likelihood solution (ML)) or by maximization of the product $P(I/O) P(O)$ (maximum a posteriori solution (MAP)). In both cases, $P(I)$ is considered as constant value which has no effect on the maximization process, and is ignored. In this class, the Lucy-Richardson algorithm [16, 18, 19] is an iterative technique used heavily for the restoration based on Bayes approach. It is widely used in the restoration of astronomical images [25].

Blind deconvolution methods represent the last class of algorithms. The performances of the preceding methods are based on a good knowledge of the PSF and also of the noise [12, 24]. The blind methods are more proper when the PSF and noises in the image are both unknown. Many of these algorithms derived from probabilistic modes compute and use the same approach as constrained iteration algorithms, including the both estimations of the PSF and the original image in each iteration. In these methods, the estimated PSF can be injected to increase the convergence of calculations.

In addition to these four classes, methods based on the wavelets seem to give good results [25]. In this work, we tested only the methods present in the four classes.

2.4 Performance Measurement

In our work, two tests are used to measure the performances of restorations methods. First, the restored results are compared at each iteration to the original image using MSE. The MSE measures the difference in energy between the two compared images. The MSE is given as:

$$MSE = \frac{1}{MN} \sum_{i,j} |O_{i,j} - \hat{O}_{i,j}|^2, \quad (6)$$

But this measure is global and it is sometime difficult to choose the best method among many approaches. So, in order to completely assess the success of our restoration procedures, we use a specific evaluation criteria based on that proposed in [14] and used by Mignotte and Meunier for SPECT images [17]. We define regions of interest (ROI) which are the objects in the test image with a high gray level. In these specific areas, we compute the mean of the gray levels to compare it to the grey level of the same ROI in the original image. Also, standard deviation of this mean is calculated. For a good restoration, the mean value must tend to the original value and the standard deviation must decreases towards zero.

3 Experiments and Results

3.1 PSF of the System

Since the application concernes the subcutaneous tumors, we record the PSF of the acquisition system including the skin of mouse in order to take into account the phenomena of absorption and diffusion of this part of the animal. To do this, we realized an acquisition system which is composed of a point source obtained using optical fiber whose diameter of heart is equal to 200 μm , lower than the size of a pixel of the image (size of the pixel being of 300 μm). To remain in the spectral band of the luciferase *in vivo*, the optical fiber is supplied by a white light source filtering by two interferometric filters. Then, to reduce the effects of the noise, obtained images represents an average of twenty acquisitions carried out under the same conditions.

3.2 Generation of Test Images

To test the performance of the restoration algorithms, we generated two test images integrating the imperfections of the acquisition system according to the degradation model expressed by the equation 1. The images are realized as a convolution between the undistorted image of shapes and the experimental PSF with adding the Poisson noise affect. Then, Gaussian noise of the system is added. Fig. 2 represents the two test images before and after degradation. The first image is formed by a disc of ray



Fig. 2. (a) (b) Test image before degradation (c) (d) Test image after degradation, respectively

2.25 mm with a uniform intensity of 6.10^4 and the second by two small close discs of ray 900 μm separated by 300 μm with a uniform intensity of 45.10^3 .

3.3 Restoration of the Test Images

First, we have tested the deconvolution methods on test images. The advantages of working with synthetic degraded images are that we know exactly the initial object, and the degradation model. Then, it is easy to quantify results by comparing the restored image to the original image (Fig. 2a and Fig. 2b). To quantify the general quality of restoration, we compute the MSE between the restored image and initial undegraded image for each method and for each iteration see Fig. 3. We report on the Tab. 1 two objects after restoration with the minimum measure of the MSE. In first consideration, we observe that statistical methods and the Tikhonov Miller method with a good choice of parameters give the best results. But the level of gray under the shapes presents ringing artifacts.

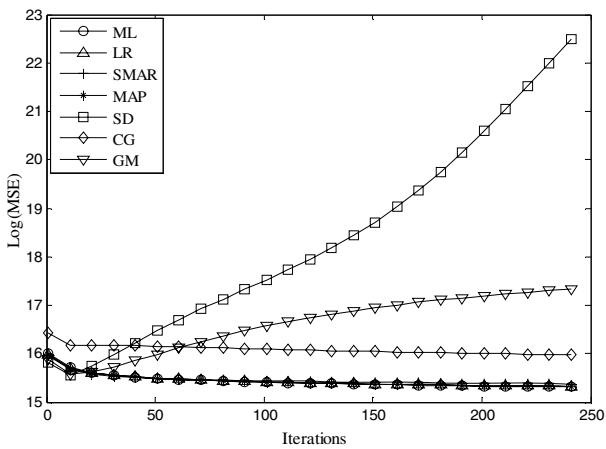


Fig. 3. The MSE of restoration methods as function of iterations number

The second experiment tests the ability of the methods to improve the quantitative measurement of the total amount of bioluminescence inside the disc area (ROI). The Fig. 4a and Fig. 4b show the evolution of the two parameters (mean and standard

Table 1. Restoration results of the test images by the different methods, with the minimal MSE. W: wiener, TM: Tikhonov Miller, GM: Golden Meinel, SD: Steepest Decent, CG: Conjugate Gradient, LR: Lucy Richardson, MAP: Maximum a posteriori, SMAR: Simultaneous Multiplicative Algebraic Restoration, ML: maximum Likelihood.

Tested Image	Linear Methods		Nonlinear Methods			Statistical Methods			Blind Methods
	W	TM	GM	SD	CG	LR	MAP	SMAR	
MSE (10^6)	7.312	4.797	5.830	6.008	15.634	4.514	4.5157	4.732	4.441
N° Iterations	-	-	20	20	250	250	250	250	250

deviation) as function of iteration number. It can be seen from Fig. 4a that the mean value for all methods converges towards the real value ($6 \cdot 10^4$) with a small overestimation except for gradient conjugate which over-estimates too much this value. For the standard deviation the statistical methods have small gap and continue to decrease with iterations as shown in Fig. 4b. On the other hand the nonlinear methods diverge after some iterations. To obtain a quantitative comparison, we take the two parameters with their optimal values i.e. the value of the number of iterations corresponding to minimum of the MSE. The results are summarized in Tab. 2. With our criteria exposed in section 2.4 and the results presented in Tab. 1 and Tab. 2, it can be easily seen that statistical restoration methods give good results.

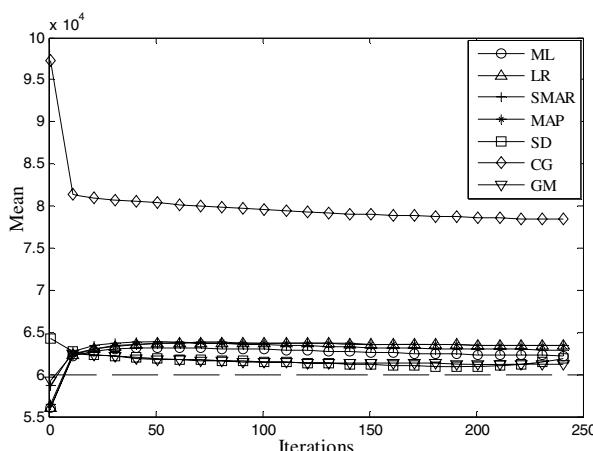
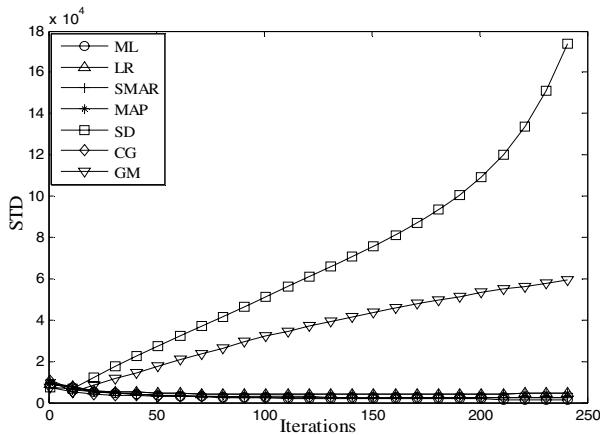


Fig. 4a. Mean of the ROI at each iteration

**Fig. 4b.** Standard deviation of ROI at each iteration**Table 2.** Mean and standard deviation of the ROI (inside the disk area) after deconvolution for each method

	Linear Methods		Nonlinear Methods			Statistical Methods			Blind Methods
	W	TM	GM	SD	CG	LR	MAP	SMAR	ML
Mean (10^4)	6.167	6.207	6.239	6.207	8.133	6.335	6.337	6.289	6.223
STD (10^3)	15.322	7.279	5.936	8.604	22.066	4.361	4.383	2.508	1.558
N° Iterations	-	-	20	20	250	250	250	250	250

3.4 Bioluminescence Images Restoration

We have also tested the deconvolution algorithms on real bioluminescence images. We present two cases where the ROI of the real acquisitions are reported in the

Table 3. Restoration of real bioluminescence images with the same condition like tests images

	Linear Methods		Nonlinear Methods			Statistical Methods			Blind Methods
Real image	W	TM	GM	SD	CG	LR	MAP	SMAR	ML

Tab. 3. The obtained results are shown in Tab. 3 and we can see that statistical methods give restored images with limiting ringing artifacts and good spatial resolution. Note that it is difficult to define a quantitative measure of the deconvolution results because we do not know the original image.

4 Conclusions

In this work nine algorithms for restoration methods were tested and compared on test images and on real bioluminescence images. To do this, we have defined a physical model of bioluminescence imaging formation. Then, we used an experimental approach to estimate more accuracy the point spread function of the acquisition system including the skin of mice. These allowed us to simulate synthetic data. After having defined and used comparison criteria, we conclude that statistical methods are more appropriated to bioluminescence imaging. On real data, we have noticed that the statistical restoration methods improve the contrast and spatial resolution of these images. To confirm this study, we will carry out physical measurements on small tumors and compare them to measures extracted from restored images.

References

1. Beck, J.A., Lloy, S., Hafezparast, M., et al.: Genealogies of Mouse Inbred Strains. *Nature Genetics* 24, 23–25 (2000)
2. Contag, C.H., Jenkins, D., Contag, P.R., Negrin, R.S.: Use of Reporter Genes for Optical Measurements of Neoplastic Disease In Vivo. *Neoplasia* 2, 41–52 (2002)
3. Edinger, M., Cao, Y., et al.: Advancing Animal Models of Neoplasia through in Vivo Bioluminescence Imaging. *European Journal of Cancer* 38(16), 2128–2136 (2002)
4. Greer, L.F., Szalay, A.A.: Imaging of Light Emission from the Expression of Luciferases in Living Cells and Organisms: A Review. *Luminescence* 17(1), 43–74 (2002)
5. Roda, A., Pasini, P., Mirasoli, M., Michelini, E., Guardigli, M.: Biotechnological Applications of Bioluminescence and Chemiluminescence. *Trends in Biotechnology* 22(6) (2004)
6. Rice, B.W., Cable, M.D., Nelson, M.B.: In Vivo Imaging of Light-Emitting Probes. *Journal of Biomedical Optics* 6(4), 432–440 (2001)
7. Sadikot, R.T., Blackwell, T.S.: Bioluminescence Imaging. *Proceedings of the American Thoracic Society* 2, 537–540 (2005)
8. Rehemtulla, A., Stegman, L.D., et al.: Rapid and Quantitative Assessment of Cancer Treatment Response Using In Vivo Bioluminescence Imaging. *Neoplasia* 2(6), 491–495 (2000)
9. Snyder, D.L., Hammoud, A.M., White, R.L.: Image Recovery from Data Acquired with a Charge Coupled Device Camera. *Journal of Optical Society of America* 10(5), 1014–1023 (1993)
10. Banham, M.R., Aggelos, A., Kastaggelos, K.: Digital Image Restoration. *IEEE Signal Processing Magazine* 14(2), 24–41 (1997)
11. Prakash Dhawan, A., et al.: Image Restoration by Wiener Deconvolution in Limited-View Computed Tomography. *Applied Optics, Journal of Optical Society of America* 24(23) (1985)
12. Kundur, D., Hatzinakos, D.: Blind Image Deconvolution Revisited. *IEEE Transactions on Signal Magazine* 13(13), 61–63 (1996)

13. Van Kempen, G.M.P., et al.: A Quantitative Comparison of Image Restoration Methods for Confocal Microscopy. *Journal of Microscopy* 185(3), 354–365 (1997)
14. Webb, S., Long, A., Ott, R.J., Leach, M.O., Flower, M.A.: Constrained Deconvolution of SPECT Liver Tomograms by Direct Digital Image Restoration. *Medical Physics* 12(1), 53–58 (1985)
15. Biemond, J., Lagendijk, R.L., Merseau, R.M.: Iterative Methods for Image Deblurring. *Proceedings of the IEEE* 78(5), 856–883 (1990)
16. Richardson, W.H.: Bayesian-Based Iterative Method of Image Restoration. *Journal Optical Society of America* 62(1), 55–59 (1972)
17. Mignotte, M., Meunier, J.: Three-Dimensional Blind Deconvolution of SPECT Images. *IEEE Transactions on Biomedical Engineering* 47(2), 274–280 (2000)
18. Shepp, L.A., Vardi, Y.: Maximum Likelihood reconstruction for emission tomography. *IEEE Transactions on Medical Imaging* 1, 113–122 (1982)
19. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data Via the EM Algorithm (With Discussion). *Journal of the Royal Statistical Society* 39(1), 1–38 (1977)
20. Lantéri, H., Theys, C.: Restoration of Astrophysical Images - The Case of Poisson Data with Additive Gaussian Noise. *EURASIP Journal on Applied Signal Processing* 2005(15), 2500–2513 (2005)
21. Katsaggelos, A.K.: Iterative Image restoration Algorithms. *Optical engineering* 28(7), 735–748 (1989)
22. Rooms, F., et al.: PSF estimation with application in auto focus and image restoration. In: *IEEE Signal Processing Symposium*, Leuven, Belgium (2002)
23. Sarder, P., Nehorai, A.: Deconvolution Methods for 3-D Fluorescence Microscopy Images. *IEEE Signal Processing Magazine* 23(3), 32–45 (2006)
24. Du, H., Voss, K.J.: Effects of Point-Spread Function on Calibration and Radiometric Accuracy of CCD Camera. *Journal of Optical Society of America* 43(3), 665–670 (2004)
25. Starck, J.L., Pantin, E., Mertagh, F.: Deconvolution in Astronomy: a Review. *Astronomical Society of the Pacific* 114, 1051–1069 (2002)

Texture-Based Multiscale Segmentation: Application to Stromal Compartment Characterization on Ovarian Carcinoma Virtual Slides

Nicolas Signolle^{1,2}, Benoît Plancoulaine², Paulette Herlin²,
and Marinette Revenu¹

¹ GREYC ENSICAEN UMR 6072 6, Bvd Maréchal Juin 14050 CAEN France

² GRECAN EA 1772, iFR 146 ICORE, Université de Caen, CLCC François Baclesse
2, Avenue du Général Harris 14076 CAEN cedex 5 France

nicolas.signolle@greyc.ensicaen.fr

Abstract. A multiscale segmentation strategy using wavelet-domain hidden Markov tree model and pairwise classifiers selection is tested in the present paper for histopathology virtual slide analysis. The classifiers selection is based on a study of the influence of hyper-parameters of the method. Combination of outputs of selected classifiers is then done with majority vote. The results of the segmentation of various types of stroma of ovarian carcinomas are presented and discussed.

Introduction

It is often necessary to identify and quantify cell and tissue compartments on histological sections, to estimate the potential evolution of the cancer lesions. To perform this measure, in an objective way, an image of the whole preparation (so-called *Virtual Slide*, *VS*) must be recorded and fully analyzed, as structures of interest are often heterogeneously spread all over the slide. It is also mandatory to adapt the working resolution to the size of the structures to be measured. High resolution *VS* produced for the identification of tiny microscopical structures occupy a very large volume in memory (several GigaBytes) and cannot be processed at once. Up to now, works performed on these large size images are scarce [12]. The method chosen by our team is a multiscale analysis of a high resolution *VS*, allowing to adapt the working resolution to the various structures to be segmented (cancer lesion, intra tumoral tissue compartments, cells and intra cellular structures for example).

The present study aims first at partitioning cancer cells and intra tumoral connective tissue and then at differentiating the various stromal compartments (fibrous tissue, loose mesenchymatous connective tissue, inflammatory cell accumulation). Stromal compartments can be identified mainly by their cell shape and organization. Then, each stromal component must exhibit an individual pattern which has to be characterized by texture analysis.

Many studies have been conducted on the segmentation of textured images. One can distinguish three categories of methods: the statistical methods which exploit the spatial distribution of the grey levels, the methods based on the construction of a texture model, and the methods which mimic the mechanism of the human vision, integrating several levels of resolution [3].

We chose to use the *Hidden Markov Tree* (HMT) proposed by Crouse [4] that combines these different approaches. It allows a statistical modeling of intra-scale and interscale properties of coefficients obtained by wavelet transform (*WT*) [5], and it exploits the decorrelation power of this transform. The goal is to capture the interscale dependency factor and the non-gaussian distribution of the coefficients computed at each scale. This model is applied to the segmentation of the various compartments of stroma identified inside VS of ovarian carcinomas, recorded at a high resolution ($0.5 \mu\text{m}$).

The paper is organized as follows : because of the complexity and variability of the structures to be segmented, the “domain objects” are first precisely described, the mathematical background of the used tools is then detailed, the results are presented and discussed before concluding on the offered prospects.

1 Material

When a patient undergoes surgery for tumour resection, the excised biopsy is fixed in formalin, paraffin embedded and splitted off in $5 \mu\text{m}$ thick slices. Slices are then affixed on a glass slide prior to be stained by imunohistochemistry or histochemistry. The special staining aims at revealing particular cell types or proteins which amount can then be estimated in order to assess the potential evolution of the tumour and to help choosing the best therapy protocol. In the case of ovarian carcinoma [Fig.1] the prognostic impact of the proportion of stroma is demonstrated [6]. The researchers want to further clarify the participation of the various stromal compartments in the evolution of tumours and their response to therapy.

The ovarian cancer lesion is made of two types of tissue: cancer epithelial cells and stroma. Stroma refers to the connective tissue which provides supportive framework and nutriments to epithelial cancer cells.

After hematoxylin staining, several types of stromal tissue can be identified inside tumours, some of which corresponding to various maturation degrees: loose connective tissue, cellular stroma and fibrous connective tissue. Inflammatory cell riched foci can also be found [Fig.2].

- The loose connective tissue is an immature stroma. It is made of an abundant extracellular matrix and star shaped mesenchymatous cells. The density of these cells is low. Cell nuclei are small and round.
- The cellular stroma is a young connective tissue. It consists mainly of myofibroblasts organized in sheets and bundles with a high density of cells. The cell nuclei are elongated and plump.

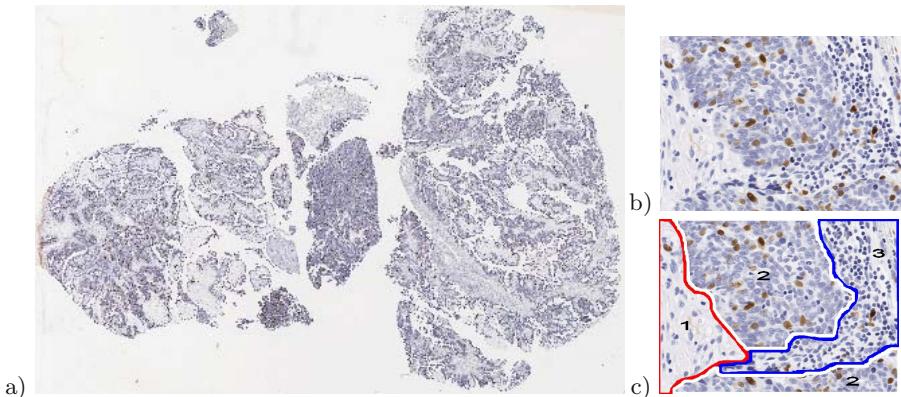


Fig. 1. a) Image of an histological section of ovarian carcinoma after DAB immunostaining of proliferation cells (brown) and Hematoxylin counterstaining of non proliferating nuclei (blue). The original size of the recorded image is 62000x41404 pixels (acquisition resolution $0.5 \mu m$). b) Small detail. c) Manual drawing of the ideal segmentation of three compartments which can be identified on this detail (1 is loose connective tissue, 2 is cancer cell foci and 3 is inflammatory compartment).

- The fibrous connective tissue is a mature stroma with a density of cells somewhat lower than the cellular stroma and a large number of extra-cellular collagen fibers. The nuclei are very thin and elongated. Their optical density is high, due to chromatin condensation.
- The inflammatory stromal compartment is characterized by small round nuclei with a high optical density (lymphocytes and plasma cells) and clover shaped dense nuclei (polymorphonuclear cells).

Epithelium appears as blue (non proliferating) and brown (proliferating) cells with large size nucleus. Extravasated red blood cells, necrosis foci and mucus sheets can also be found inside the tumor [Fig 2].

Tiled tiff VS were recorded at a resolution of $0.5 \mu m$ thanks to a high resolution slide scanner (Scanscope CS from Aperio Technologies) provided with a $20x$ objective. A 30% loss Jpeg compression was applied to each tile when recording images. Uncompressed Images are 3 to 10 GigaBytes files, depending on the scanned tissue area.

2 Methods

2.1 Multiscale Strategy

Previous works based on low resolution analysis strategy ($6.3 \mu m$) already allow to segment and quantify markers (quantification of stained nucleus, of epithelium, of blood vessels, of stroma [7,8]). However, the fine analysis of stroma compartments, which needs a cell characterization, is impossible at this scale as a cell nucleus is represented by only a few pixels.

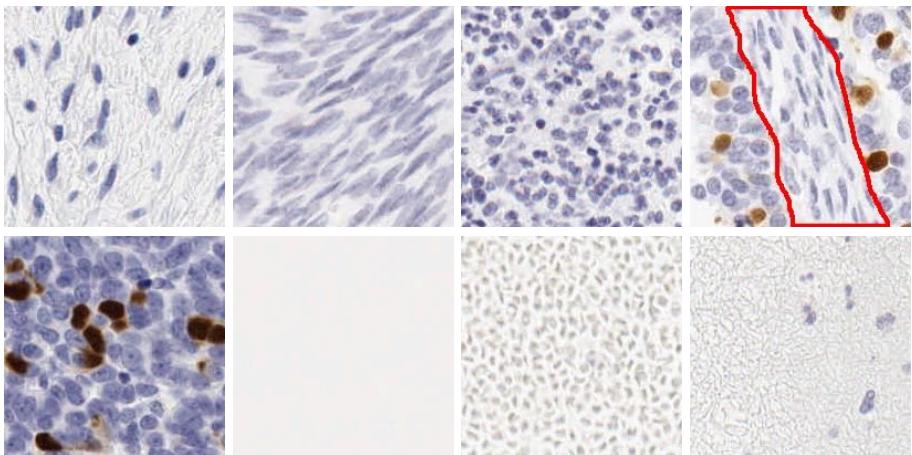


Fig. 2. First row (from the left to the right), various types of stroma: loose connective tissue, cellular stroma, inflammatory cells, and fibrous connective tissue (surrounded area). Second row: other types of compartments encountered in an image: Epithelial cancer cells, Background, Red Blood Cells, Mucus.

We propose a high resolution analysis strategy. The image's volume (3 to 10 GB) prevents to process VS at once. The tempting solution is to process tiles of a splitted image. But this implies hard constraints on elementary image edges, especially when managing large structures that can be spread over several tiles. Furthermore, large scale structure segmentation can imply local as well as global characteristics which are difficult to take into account. The proposed multiscale approach consists in selecting, at various scales, regions of interest containing objects to be further segmented. This allows the user to adjust the decomposition of the image to its contents. A large structure will be segmented at a low resolution (subsampled image size is then small enough to be held in memory, so that there will be no edge effect to manage). The result of this segmentation will serve as a mask for segmenting a smaller structure at a higher resolution. The procedure can be reiterated until localization of tiny structures inside the full resolution virtual slide.

We used wavelet transform [5] to subsample the images. A first partitioning between background and tissue can be done at a very low resolution ($32 \mu m$). For this purpose, the image histogram was regarded as a mixture of three gaussian curves (one mode for background, and two closer modes for stroma and epithelium). Parameters of these three gaussian curves can be estimated by an histogram analysis. The likelihood of a pixel to belong to a class (background or tissue) can be computed and the image can then be segmented by maximum of likelihood. The resulting binary mask of tissue compartment was improved thanks to mathematical morphology operations (hole fill and geodesic opening).

The HMT model was applied to tissue compartment using marginal statistics of wavelets coefficients.

2.2 Hidden Markov Tree (HMT) Model

At each scale, the value of coefficients of a wavelet transform depends on the regularity of the signal. A singularity would yield a large coefficient of wavelets that could cascade through scales, while an area with small variation values would yield a series of small coefficients. These properties of the wavelet transform mean that there is a small number of large coefficients bringing the mainpart of the energy of the signal, and many small coefficients. To approximate both the marginal and joint wavelet coefficients statistics, Crouse [4] proposed to use a model of Gaussian mixture applied to wavelet coefficients in each sub-band, and a hidden Markov tree (HMT). The HMT model associates to each wavelet coefficient a (hidden) state variable that controls whether it is Large (L) or Small (S). The joint density of each coefficient is then modeled as a gaussian mixture with a high variance and null average gaussian for the L state and with a low variance one for the S state. The model captures inter-scale dependencies between the wavelet coefficients using a Markov chain, with one dependency between the resolution levels [Fig 3].

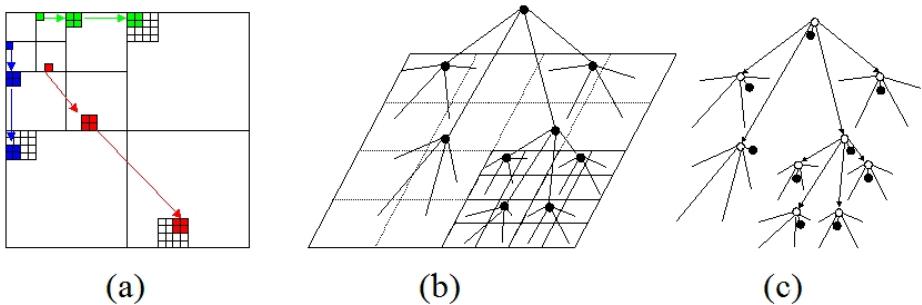


Fig. 3. Drawing from [9] a) The father-son dependencies of WT sub-bands. (b) The quad-tree structure of the model, detailed for a sub-band. (c) The HMT model, detailed for a sub-band. Each wavelet coefficient (black node) is modeled by a gaussian mixture controlled by a hidden state (white node).

This model is used by Choi [9] to achieve the segmentation of textured images. The method consists in three phases: a learning phase, a phase of segmentation at each level and a phase of inter-scales merging. During the learning phase, the model parameters \mathcal{M} are learnt for each texture, through an EM (*Expectation Maximization*) algorithm suited to HMT [4]: for each state (L or S), each level and each sub-band, the parameters of the representative Gaussian shapes (mean, variance and probability of occurrence) are learnt, as well as the parameters of transition from one state to another, between two resolutions levels:

$$\begin{bmatrix} \varepsilon_{s=S}^{f=S} & \varepsilon_{s=L}^{f=S} \\ \varepsilon_{s=S}^{f=L} & \varepsilon_{s=L}^{f=L} \end{bmatrix} = \begin{bmatrix} \varepsilon_{s=S}^{f=S} & 1 - \varepsilon_{s=S}^{f=S} \\ 1 - \varepsilon_{s=L}^{f=L} & \varepsilon_{s=L}^{f=L} \end{bmatrix} \text{ where } \varepsilon_{s=b}^{f=a} \text{ represents the probability}$$

that the father is in state a while the son is in state b .

During the phase of segmentation, the likeness between the data of the wavelets transform of the observed image and the HMT model of learnt images is assessed by calculating a likelihood function :

$$f(\mathcal{T}_i | \mathcal{M}) = \prod_{\text{orientations}} \sum_{m=S,L} f(\mathcal{T}_i | S_i = m, \mathcal{M}_o) \times p(S_i = m | \mathcal{M}_o)$$

where \mathcal{T}_i is a sub-tree of root w_i in one of the sub-bands, \mathcal{M}_o represents the model parameters for a particular orientation and S_i is the hidden state of w_i . Each pixel is then affected, with the *Maximum of Likelihood* method, to the more similar texture. This yields to a “raw” segmentation at each level.

To improve this first segmentation, the segmentations at different resolutions are subsequently merged. This merging is achieved by maximizing the likelihood of a pixel to belong to a class, taking into account both its value in the raw segmentation and a context vector. This context vector can correspond to the value of the father and of the neighbors of the father [9], of the neighbors of the son, or a combination of several context vectors [10] [Fig.4].

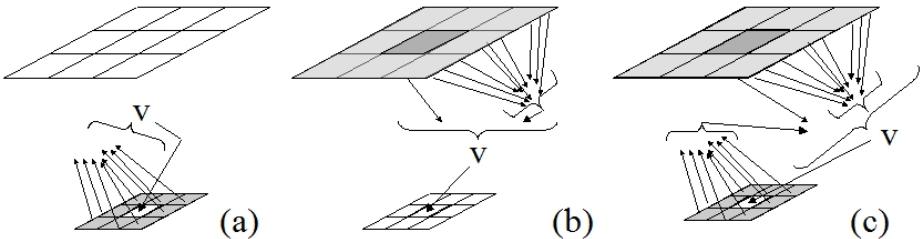


Fig. 4. Drawing from [10]. The context vectors V used: a) neighbors of the son (b) neighbors of the father and the fathers value (c) neighbors of the father, sons and fathers value.

2.3 Selection of Hyper-parameters

The HMT parameters are calculated through a learning set of pure class images, for each combination of the method parameters. Indeed, the HMT model itself can be parameterized : one have to choose the wavelet base (and possibly its order), the color image component on which the WT is applied and the number of resolution levels on which the analysis is focused. These four hyper-parameters can greatly affect the HMT model results. Each set of hyper-parameters allows to generate a different segmentation.

Each set of hyper-parameters corresponds to a classifier, and the best of them have to be selected thanks to an evaluation of their performances. Rather than compare them for all classes together, the multiclass problem is transformed into a set of binary ones by decomposing it in one-against-one problems [11]. To evaluate the classifiers, each one is trained to become a discriminative specialist in two classes (dichotomizer). For this purpose, a base containing images with

two classes solely is used. HMT segmentation is realized by proposing exclusively these two classes as a possible result of segmentation. The classifier kept for distinguishing the two classes, is the one that maximizes the proportion of well segmented pixels on the two classes compared to the manual drawn “ground truth”. The proportion of well segmented pixels, for each class, represents the confidence attached to the classifier thereafter. There is so $N = n * (n - 1)/2$ binary classifiers (where n is the number of classes). Each classifier uses parameters learned by the HMT model for one set of hyper-parameters.

2.4 Merging of Each Classifier Segmentation

To merge the outputs of the N classifiers, many combination rules can be applied to couple the estimates of each binary classifier in order to obtain class membership estimates for the multiclass problem.

The most commonly used combination rule is probably the *Majority Vote* [12]. Class membership estimates are computed according to the formula:

$$p(\omega = \omega_i | x) = \frac{1}{N} \sum_{j=0, j \neq i}^N I(\Psi_{\omega_i, \omega_j} | x) \times Cf(\omega_i, \Psi_{\omega_i, \omega_j}) \text{ where } \Psi_{\omega_i, \omega_j} | x \text{ is the probability}$$

of segmentation result (class ω_i or ω_j) with classifier $\Psi_{\omega_i, \omega_j}$ at pixel x . $I(\Psi_{\omega_i, \omega_j} | x) = 1$ if $\Psi_{\omega_i, \omega_j} | x = \omega_i$ and 0 otherwise. $Cf(\omega_i, \Psi_{\omega_i, \omega_j})$ is the confidence in $\Psi_{\omega_i, \omega_j}$ for class ω_i . The resulting label is computed as $\text{argmax}(p(\omega | x))$.

One can also use combination rule based on *Error Correcting Output Codes* (*Ecoc*) [13]. Each class is represented by its own output code in the output vector space. The chosen class is the one that is the closest to the 0/1 prediction vector obtained from the classifiers. The *Loss Based Decoding* (*LBD*) [14] is an improvement of *Ecoc* because it allows to use directly the outputs of the binary classifiers instead of their hard 0/1 predictions.

3 Experimentation and Results

The tests have been conducted with several wavelets bases (Haar, Battle Lemarie, Daubechies and bi-orthogonal Splines). Two orders of wavelets (a low one and a high one) were used when possible and one to five levels of HMT decimation (resulting in four levels of segmentation) were computed. WT was computed on red, green, blue and luminance components as well as on excess red image (twice red minus green and blue), on excess green image and on excess blue image. Scheunders representation was also used. Scheunders [15] proposed a method to construct a wavelet representation combining the color plans calculating, based on the three values, a coefficient reflecting their relative importance. These parameters constitute a total of 192 different segmentations. HMT training set consists of five 512x512 images of each class. Classifiers are selected on an other training set of four 1024x1024 images per pair of class (*i.e.* 40 images).

The training set was improved several times, to take into account some variability within the same class, until its images are well segmented by the process.

The resulting hyper-parameters after selection are summarized in Table 1.

Table 1. Parameters allowing good discrimination of compartments. (Epi stands for cancer epithelial cells, CS for cellular stroma, Back for background, LCT for loose connective tissue and IS for inflammatory stroma).

Compartment 1	Compartment 2	Wavelets Base	Wavelets Order	Component	Level of decimation	Confidence in compart. 1	Confidence in compart. 2
Epi	Back	Haar	/	Scheunders	1	99.81	99.84
Epi	LCT	Haar	/	Scheunders	2	98.91	93.32
Epi	CS	Splines	5	Xred	4	81.88	85.76
Epi	IS	Battle	1	Xgreen	3	98.25	92.83
Back	LCT	Battle	1	red	2	98.95	99.65
Back	CS	Haar	/	Scheunders	1	99.80	99.57
Back	IS	Haar	/	blue	2	99.07	99.53
LCT	CS	Haar	/	red	2	94.07	98.13
LCT	IS	Daubechies	10	Xgreen	4	96.30	98.98
CS	IS	Daubechies	2	Xgreen	3	97.48	90.56

Various combination rules were tested to merge the binary classifiers : *Error Correcting Output Codes* , *Loss Based Decoding*, but they did not provide a significant improvement over Majority Vote.

The testing set is composed of fifteen 2048x2048 images. Compared to reference manual segmentation, 60.05 to 89.67 percent of well labeled pixels are obtained together with correctly localized borders [Fig.5]. Average is 71.02%. The segmentation result has been improved with further spatial regularization to ban tiny area. The main default is that in some cases, epithelium may be confused with cellular and inflammatory stroma (the closest compartments).

Tests are conducted on one processor of a 3GHz Xeon quadriprocessor. Processing time is from ten minutes to two hours per image. Mean time is half an hour. That is about one hundred to three hundred hours for the whole VS, depending of the amount of background.

4 Conclusion and Prospects

In this paper, we have investigated a strategy for segmenting various stromal compartments on ovarian carcinoma virtual slides. We proposed to use a wavelet domain hidden Markov tree model to get classifiers. Selection among these classifiers is done with a *one against one* approach. Classifiers outputs combination was performed thanks to a *Majority Vote* rule. Promising results are obtained on few selected test images and should be confirmed on a larger test set. The ratio of well classified pixels obtained must be appraised taking into account the fact that the transitions between regions of stroma are not straightforward and sometimes difficult to identify even for an expert. However, some mislabeling of the epithelium may happen, and processing time (about one hour for a 2048x2048 image) due to *Expectation Maximization* algorithm's convergence is prohibitive for any clinical use.

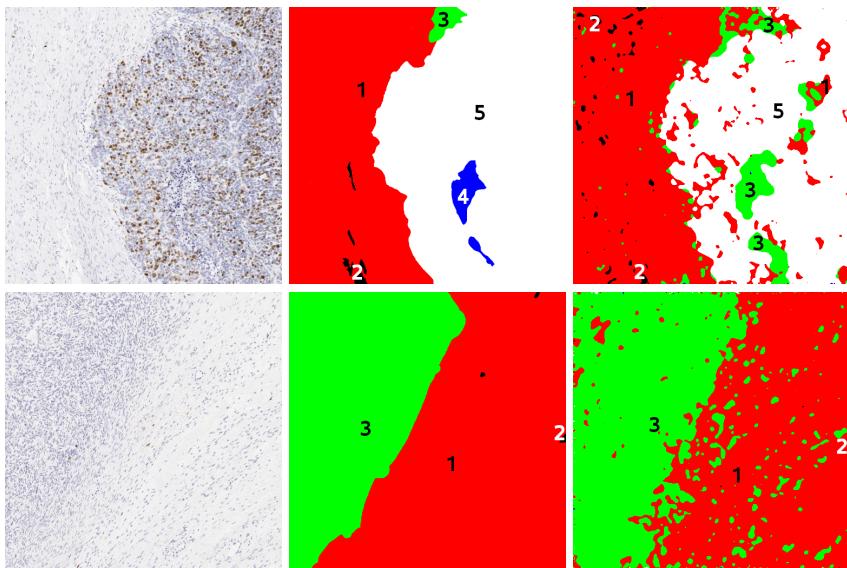


Fig. 5. Left : image to be segmented, middle : manual segmentation, right : automated segmentation. 1 is LCT, 2 is Back, 3 is CS, 4 is IS, and 5 is Epi.

To answer the problem of cancer cells foci detection, we will use previously described methods [8] to locate epithelium at low resolution and eliminate it from the higher resolution analysis.

To cope with the processing time, we have to optimize our code and consider replacing EM part with a faster converging method. We also plan to study other segmentation approaches including image patch exemplar based methods.

Acknowledgments

This work was funded by a BDI grant cofinanced by the CNRS and the Regional Council of Lower Normandy.

References

1. Wang, Y., Turner, R., Crookes, D., et al.: Investigation of methodologies for the segmentation of squamous epithelium from cervical histological virtual slides. In: IMVIP 2007, Int. Machine Vision and Image Processing Conference, pp. 83–90 (2007)
2. Pan, T.C., Mosaliganti, K., Machiraju, R., Cowden, D., Saltz, J.H.: Large scale image analysis and nuclear segmentation for digital microscopy images. In: APIII, Frontiers in Oncology and Pathology Informatics, Pittsburgh (2004)
3. Dombre, J.: Systèmes de représentation multi-échelles pour l’indexation et la restauration d’archives médiévales couleur. PhD thesis, University of Poitiers, France (2003)

4. Crouse, M.S., Nowak, R., Baraniuk, R.: Wavelet-based statistical signal processing using hidden markov models. *IEEE Trans. Signal Processing* 46, 886–902 (1998)
5. Mallat, S.: A wavelet tour of signal processing. Academic Press, London (1999)
6. Elie, N., Labiche, A., Michels, J.J., Herlin, P.: Control of low resolution scanning of ovarian tumor stromal compartment. *Image Analysis and Stereology* 24(2), 85–93 (2005)
7. Tran, T., Elie, N., Plancoulaine, B., et al.: An original approach for quantification of blood vessels on the whole tumour section. *Analytical cellular pathology* 25(2), 63–75 (2003)
8. Elie, N., Plancoulaine, B., Signolle, J.P., Herlin, P.: Simple way of quantifying immunostained cell nuclei on the whole histologic section. *Cytometry* 56A(1), 37–45 (2003)
9. Choi, H., Baraniuk, R.: Multiscale image segmentation using wavelet-domain hidden markov models. *IEEE Trans. on Image Processing* 10(9), 1309–1321 (2001)
10. Fan, G., Xia, X.G.: A joint multicontext and multiscale approach to bayesian segmentation. *IEEE transactions on geoscience and Remote Sensing* 39(12), 2680–2688 (2001)
11. Hastie, T., Tibshirani, R.: Classification by pairwise coupling. *Annals of statistics* 26(2), 451–471 (1998)
12. Friedman, J.H.: Another approach to polychotomous classification. Technical report, Department of Statistics, Stanford University (1996)
13. Dietterich, T.G., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research* 2, 263–286 (1995)
14. Allwein, E.L., Schapire, R.E., Singer, Y.: Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research* 1, 113–141 (2000)
15. Scheunders, P.: An orthogonal wavelet representation of multivalued images. *IEEE Transactions Image Processing* 12(6), 718–725 (2003)

Combination of Image Registration Algorithms for Patient Alignement in Proton Beam Therapy

Rachid Belaroussi and Guillaume Morel

Institut des Systèmes Intelligents et Robotique,
4 place Jussieu, 75005 Paris, France
{belaroussi,morel}@robot.jussieu.fr
<http://www.isir.fr>

Abstract. We propose a measure of patient alignment in a video by combining different image representations : grey level, edges, and a set of feature points. When patient head is correctly positionned, a reference image with its ellipse is stored as a template of correct alignment. Edges detection results in a second template of the correct head location. Corners inside the ellipse are detected and tracked: a set of N feature points composes a third template. Template matching computes a measure of similarity between a representation of the reference image and a window sliding around the reference location. Similarity with these three models are combined by the product rule. Location of window the most similar to the templates gives the translation \mathbf{T} of the reference model in the image plane. This measure of patient misalignment could avoid X-ray verification of patient alignment, reducing patient dose and duration of treatment sessions.

Keywords: Proton beam therapy, expert combination, template matching, feature points, color model, camshift.

1 Introduction

Conformal radiotherapy is a recent approach for tumor removal that uses an aperture, a metal block with a hole through which the radiation beam passes. Each patient has a custom-made aperture, the shape of the hole is the approximate shape of the target being treated by the beam. Protontherapy aims energetic ionizing particles (protons) onto the target tumor: protons scatter less easily in the tissue than X-rays and the beam stays focused on the tumor shape without much lateral damage to surrounding tissue. Also, no proton penetrates beyond a depth corresponding to the Bragg peak: dosage to tissue is maximum in the tumor volume and almost null outside. A slight error in tumor positioning can therefore damage surrounding tissue: the required accuracy in patient positionning in the beam line is only 1 mm.

The patient positioning system is made of a patient couch moved by a robot: patient alignement refers to verification of the patient positionning. Before and during a treatment session, patient head is kept still on the couch by a thermoplastic

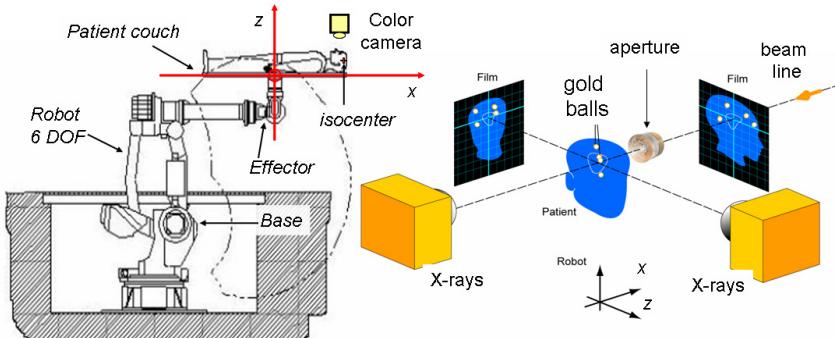


Fig. 1. Manipulator robot and patient couch: robot arm places a load from 15 kg to 150 kg at 3,50m from its base. Correctness of patient head alignment is validated using two orthogonal x-rays images to detect fiducial markers implanted in skull.

head mould. To ensure the brain tumor is well positioned in beam line a patient alignment is required before the treatment starts, as well as a patient monitoring during the treatment to stop the proton beam when patient head is moving.

Our goal is to develop a measure of positioning error of patient head, with a color camera. Our approach is non invasive, no marker of the tumour is visible so the object tracked is the head mould, which is a non rigid object. This task requires an approach accurate enough to measure small displacement ≤ 1 mm, robust to lighting conditions and fast. Our perspective is a visual servoing of a 6 degree of freedom positioner robot : the difference with the template is corrected by a displacement of the robot.

A X-rays alignment procedure is used for patient setup verification : two X-rays images of orthogonal sections of the head are taken to validate the positioning as shown by figure 12. Fiducial markers (golden balls) implanted in the patient skull are detected by an operator, and a correction is sent to the positionner. The positionning is X-ray checked again until convergence. A similar approach of correspondance between current fiducial markers orthogonal images and reference is found in [12], and can be applied in radio-surgery [13]. A refinement step can be provided by an infrared stereovision camera POLARIS that returns 3D locations of five reflecting ball implanted on the plastic mould [6]. Other patient alignment systems use image registration techniques based on skull image, or external markers detected in 3D with several infrared stereovision cameras [9], [8]. For the treatment of mobile tumour in soft tissues like liver, such as infrared stereovision markers are used to monitor respiration phase, with electromagnetic markers implanted close to the tumour [11]. A model of tumour mobility versus surface displacement is built to compensate motion due to respiration phase. AlignRT [10] proposes a 3D surface reconstruction of a part of the body, using a structured lighting and a stereovision camera. It is used in patient respiration phase monitoring to command a 4 dof robot, and is tested in proton therapy for brain tumour at Massachusetts General Hospital. For prostate radiotherapy

targeting, a comparative study between ultrasound and x-ray imaging of markers is proposed in [7].

Our work aims at replacing that last validation step by a visual servoing of the patient positioning system, using a color camera. In this study, we investigate the use of template matching over different image representations to estimate the translation $\mathbf{T} = [T_x \ T_y \ T_z]^T$ vector required to align the current couch position to the correct patient positioning.

Section 2 presents our procedure of patient alignment, head mould being modeled as an ellipse used for template building and matching. Section 3 describes models based on various representations of head mould (grey vel and image gradient), and section 4 experts combination is formulated. Experimental results are discussed in section 5, and section 6 gives conclusion and perspectives of this work.

2 Patient Alignment System Overview

In a first session, an image I_{grey}^* of patient head is acquired when the patient positioning is known as correct: head location is represented by an ellipse with the state \mathbf{s}^* , estimated by CamShift. The template matching is implemented in a region close to the location of reference ellipse \mathbf{s}^* : the positioner translation error is supposed to be less than 1 cm.

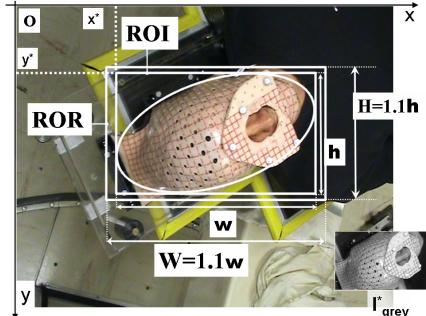


Fig. 2. Region of interest (ROI) is the rectangle bounding the template ellipse: region of research ROR of the matching window is set around the ROI

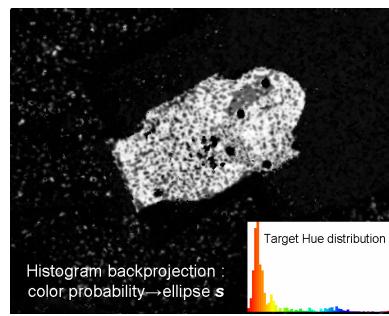


Fig. 3. Backprojection of head mould color model (a 64 bins histogram in the Hue channel of HSV): second moment order are computed to estimate ellipse

The initialisation stage stores the template image I_{grey}^* and the binary image of the ellipse ℓ^* resulting from the CamShift [4]. ℓ^* bounding box is used to define the region of interest of the frame stored in I_{grey}^* :

$$\mathbf{ROI}^* = [x_{roi}^* \ y_{roi}^* \ w \ h]$$

with (x^*, y^*) top-left corner coordinate, w is width and h is height of \mathbf{ROI}^* . A Canny edges detection of I_{grey}^* is computed: a second template is made off the

binary image I_{edges}^* of I_{grey}^* contours. All pixels of I_{edges}^* outside of ℓ^* are set to 0. A set \mathbf{E}^* of $N = 200$ 2D feature points of I_{grey}^* are detected:

$$\mathbf{E}^* = \{\mathbf{x}_k^* = (x_k^*, y_k^*)\}_{k=\{0\dots N-1\}}$$

The test stage starts with CamShift target initialisation, that can be done manually. Patient head is tracked by Camshift which results in an ellipse state estimation $\mathbf{s}(t)$ in the image at time t . Ellipse image $\ell(t)$ is used to mask pixels of edges $I_{edges}(t)$ and intensity $I_{grey}(t)$ images outside $\ell(t)$. Corners are detected at time $t = 0$ of target initialisation: N feature points are searched in $\ell(0)$ ellipse. These points are further tracked at time $t > 0$ resulting in a set of N points $\mathbf{E}(t)$:

$$\mathbf{E}(t) = \{\mathbf{x}_k(t) = (x_k(t), y_k(t))\}_{k=\{0\dots N-1\}}$$

Distance between this set and \mathbf{E}^* is used as a similarity measure. Results of template matching with the three models are combined to evaluate the translation error $\mathbf{T}(t)$. The time parameter is leaved when possible in the following sections, as the process applied to the current image except for feature points tracking.

3 Basic Models Description

3.1 Appearance Based Model

Template matching is realized by sliding a $w \times h$ window and computing the Euclidian distance between I_{grey}^* and overlapped $I_{grey}(t)$ region. It is computed on a $W \times H$ region of research \mathbf{ROR} centered on \mathbf{ROI}^* , 10% wider and larger:

$$\mathbf{ROR} = [x_{ror} \quad y_{ror} \quad W \quad H]$$

The resulting array is named **GreyMap** and have a size of $(W - w + 1) \times (H - h + 1)$, where each element correspond to a window in image at time t :

$$GreyMap(x_0, y_0) = a \sum_{x,y} (I_{grey}(x + x_0, y + y_0) - I_{grey}^*(x, y))^2 + b$$

(a,b) are used to normalize **GreyMap**: it is inversed so that a high value represents a window close to the template, and its range is linearly scaled onto $[0 \quad 1]$.

In figure 4, **GreyMap** is a 44×26 array: 1144 $w \times h$ window locations are evaluated, in the neighborhood of \mathbf{ROI}^* . To evaluate accuracy of the approach, diagonal of \mathbf{ROR} is made of about 560 pixels and represents 45 cm: image resolution is then ≈ 12 pixels per cm. Our approach is then able to estimate a translation error with an accuracy of 0.8 mm. In the case of figure 4, patient is exposed with a diagonal incidence to the beam line: translation error range is $T_x \in [0 \quad 17.6 \text{ mm}]$ along the x-axis and $T_y \in [0 \quad 10.4 \text{ mm}]$ along y-axis.

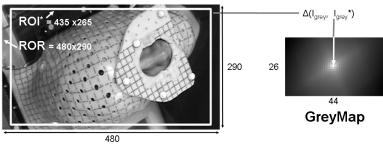


Fig. 4. Template Matching in the grey level image: GreyMap elements are correlation between reference and overlapped region

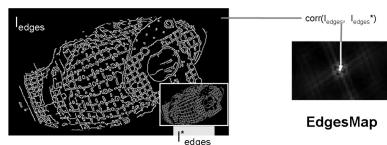


Fig. 5. EdgesMap is the result of correlation between template edges and edges image

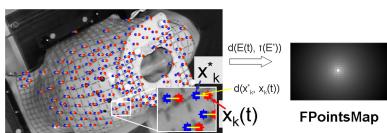


Fig. 6. FPointsMap is the template matching between reference feature points and set of points in the current image, using the Euclidian distance

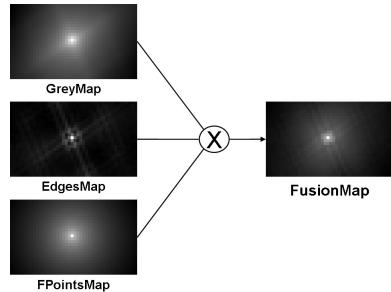


Fig. 7. FusionMap maximum correspond to window of current image the most similar with the reference models

3.2 Geometrical Model

A Canny detector is used to estimate edge pixels in the image : it results in a binary image I_{edges} . Correlation between I_{edges}^* is calculated over the search region **ROR** resulting in **EdgesMap** array, of same size as **GreyMap**. It is a classical approach of image registration [14]: assumption is made of an affine transform between intensity of the template image and a window of the region of research.

$$EdgesMap(x_0, y_0) = \sum_{x,y} I_{edges}^*(x, y) I_{edges}(x + x_0, y + y_0)$$

In figure 5, **EdgesMap** array is illustrated: pixels intensity is proportional to the correlation between the $w \times h$ block it represents and the template edges. It appears that this model is quite sensitive to translation in the image plane because of the grid drawn on the head mould. Global maxima represents the window that best matches the template, but we can also see some local maximum are regularly distributed in **EdgesMap**: they are due to the grid squares periodicity.

3.3 Structural Model

Ensemble $\mathbf{E}(t)$ of features point is detected and tracked during the sequence:

- feature points detection [3] computes correlation matrix of horizontal and vertical image gradient: pixels with the higher eigenvalues are detected.
- at time $t > 0$, optical flow of feature points is computed using iterative Lucas-Kanade [2] method in pyramids: each point $\mathbf{x}_k(t) \in \mathbf{E}(t)$ is calculated knowing $\mathbf{x}_k(t-1)$, $I_{grey}(t)$ and $I_{grey}(t-1)$.

The distance between $\mathbf{E}(t)$ and the reference set \mathbf{E}^* is computed by sliding the template over the $\mathbf{ROR}(t)$ rectangle :

$$FPointsMap(x_0, y_0) = \frac{1}{N} \sqrt{\sum_{k=0}^{N-1} ((x_k^* - x_{roi}^*) - (x_k - x_{ror}) + x_0)^2 + ((y_k^* - y_{roi}^*) - (y_k - y_{ror}) + y_0)^2}$$

where (x_0, y_0) are coordinates of the top left corner of the $w \times h$ region overlapped by the feature points template. Array **FPointsMap** size is also $(W - w + 1) \times (H - h + 1)$, but the more a window is similar to the template, the lower the distance to the template is. It is inversed and linearly scaled onto $[0 \quad 1]$ as we did for **GreyMap** array. In figure [6] **FPointsMap** array values are video-inversed: the window closest to the template is represented by a white pixel, while a darker intensity represents a window less likely to match the template \mathbf{E}^* .

4 Models Combination

4.1 Translation Error in the Horizontal Plane

We computed three $(W - w + 1) \times (H - h + 1)$ arrays storing the results of comparison between reference models and windows of the current image. In our experiments, the camera is placed above the patient couch so that the image plane is horizontal. Each of these map can be used to estimate the translation error $\mathbf{T}_{hor} = [T_x \quad T_y]^T$ in the image plane. An arbitration is necessary when the sources are in conflict: a data fusion methods is applied.

A recent study of information fusion applications and methods can be found in [15]. Goal of information fusion depends on the application : dimension reduction, classification, robustness to imperfect sources. Several architectures of data fusion are proposed in literature, and can be divided into three types: parallel, serial and hybrid. Level of data fusion refers to the degree of abstraction of inputs and outputs. At signal level, data are combined without transformation: it is the smallest abstraction level. At an intermediate level, features (dimension, area, compacity, mean ...) are extracted before combination. The highest level is the fusion of decisions or experts combination: this level is adapted to our issue.

Different methods of fusion could be used like weighted average, majority voting, maximum rule, fuzzy logic or product rule, as we only have one training sample for each image representation. The product rule is a quite efficient [4]

approach for combining experts output, it is close to intuition and more easy to interpret:

$$\text{FusionMap} = \text{GreyMap} \cdot \text{EdgesMap} \cdot \text{FPointsMap}$$

An advantage of the product rule is that the issue of comeasurability of sources is not raised. A weighted average requires normalization of physically different types of measurements : grey level, contours and feature points. A connexionist approach is not applicable: no image database is available and mould are custom-made anyway.

The product rule [5] assumes that the representations used are conditionally statistically independent. It can be a severe rule as a single expert can inhibit an interpretation by outputting a small similarity for it. However integration of new expert is simplified in this combinatory architecture; this modularity is desirable for further development. The translation error is estimated using **FusionMap** maximum location:

$$(x_{max}, y_{max}) = \max_{x,y} \text{FusionMap}(x, y) \\ \Rightarrow T_x = x_{roi}^* - (x_{ror} + x_{max}) \quad T_y = y_{roi}^* - (y_{ror} + y_{max})$$

4.2 Couch Altitude Estimation

Translation error in a vertical plane can be determined using the same approach with a camera which axis is orthogonal to the z-axis. This way, two cameras could do the same task as X-ray tubes used for patient position validation (see figure [1]), and translation $\mathbf{T}_{\text{vert}} = [T_x \quad T_z]^T$ is estimated. However, in this study our test sequences are acquired with a single camera placed above the couch: we will suppose couch altitude constant, and the z-axis translation error is not further investigated in this paper.

5 Patient Alignment Performances

As mentionned earlier, our approach is able to estimate a translation error of 0.8 mm. The computation time is **180 ms per frame** on a Pentium Celeron @1.2GHz, 240Mo RAM, with no particular optimization. To process a 720x576 image, 50 ms are required by the CamShift and 130 ms for the image processing (edges detection, pyramidal images for optical flow), template matching and fusion algorithms with a 435x265 head size.

We tested a sequence of 1500 images with a specific sceniaro: patient couch is correctly positioned, and still. Patient is also still in the beginning of the sequence, then moves his head more or less. This scenario validate our approach as a misalignement due to head motion can be detected. To quantify sensitivity of our estimate of translation error, we computed the amount of motion pixels in the intersection of ellipses ℓ^* and $\ell(t)$. Motion detection is realized by frame differencing with the template grey level image. Temporal gradient of the images sequence is computed in **ROI*** rectangle and binarized: pixels with module

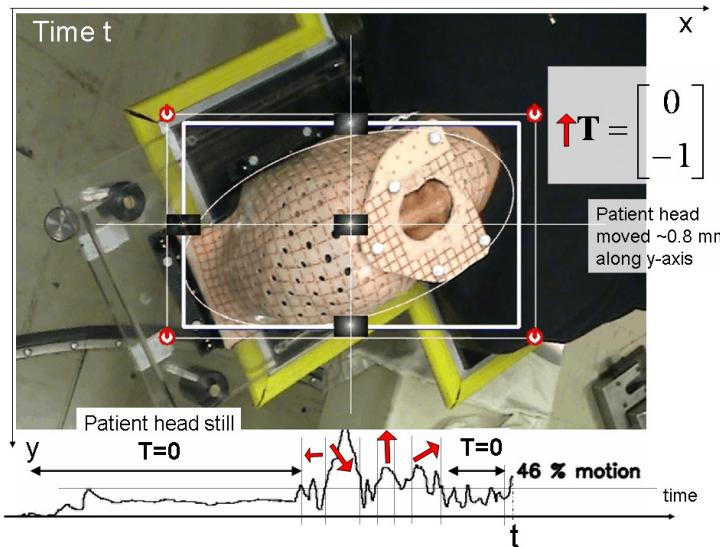


Fig. 8. Edges and feature points of templates (in blue) and in image at time t (in red): matching edges and point are drawn in white. Proportion of motion pixels in region of interest versus time during the test sequence: translation error is estimated as null when patient head is still, and is drawn with a red arrow when non zero.

higher than an experimentally defined threshold are classified as in motion in $D(t)$ image. Ratio between number of motion pixels and number of edges pixels is used to express the motion quantity in percent. When patient head moves more than 25%, a non zero translation vector is proposed by our algorithm.

In figure 8, an illustration is provided with at time t a motion of 46% and a translation error estimated of 1 pixel along the y -axis. At this moment, patient head is moving inside its mould, and the caused mould displacement is evaluated at $T_x = 0.8$ mm. It is worth noticing that when patient head is still, the translation error estimated is null, which is desirable too.

6 Conclusion and Perspectives

We presented three models of a reference based on various image representations: one based on grey level, two others based on image spatial gradient (edges and feature points). We proposed to aggregate these heterogenous sources for a measure patient alignment error in a color image. An interesting point of the approach is that, even if the grid drawn on the head mould makes the edges matching more efficient, it is not modality specific and it could be applied to patient alignment in radiotherapy or tomotherapy for example, with or without mould. Our first experiments of the proposed models fusion are very encouraging, our algorithm is fast and have a resolution of less than 1 mm. Alignment accuracy requires a ground truth to be estimated: this is to be done with an

infrared stereovision camera. Also, our next step is to implement a visual servoing of our 6 dof robot positionner based on our measurement.

References

1. Bradsky, G.: Computer Vision Face Tracking For Use in a Perceptual User Interface. *Intel Technology Journal* (1998)
2. Bouguet, J.-Y.: Pyramidal Implementation of the Lucas Kanade Feature Tracker: Description of the algorithm. Technical report, Intel Corporation Microprocessor Research Labs (2000)
3. Shi, J., Tomasi, C.: Good Features to Track. In: Conference on Computer Vision and Pattern Recognition (1994)
4. Milgram, M., Belaroussi, R., Prevost, L.: Multi-stage Combination of Geometric and Colorimetric Detectors for Eyes Localization. In: 13th International Conference Image Analysis and Processing, pp. 1010–1017 (2005)
5. Kittler, J., Hatef, M., Duin, R., Matas, J.: On Combining Classifiers. *Transactions on Pattern Analysis and Machine Intelligence* 20(3), 226–239 (1998)
6. Pinault, S., Morel, G., Ferrand, R., Auger, M., Mabit, C.: Using an external registration system for daily patient repositioning in protontherapy. In: International Conference on Intelligent Robots and Systems (2007)
7. Fuller, D.C., Thomas, C.R., Schwartz, S., Golden, N., Ting, J., Wong, A., Erdoganmus, D., Scarbrough, T.J.: Method comparison of ultrasound and kilovoltage x-ray fiducial marker imaging for prostate radiotherapy targeting. *Phys. Med. Biol.* 51(19), 4981–4993 (2006)
8. Baronia, G., Ferrigno, G., Orecchia, R., Pedotti, A.: Real-time three-dimensional motion analysis for patient positioning verification. *Radiotherapy and Oncology* 54 (2000)
9. de Kock, E.A., Muller, N., Maartens, D., van der Merwe, J., Muller, D., van Rooyen, R., van der Merwe, A., Eksteen, J., von Hoesslin, N., Wagener, D., Hough, J.: Integrating an industrial robot and multi-camera computer vision systems into a patient positioning system for high-precision radiotherapy. In: ISR Symposium (2004)
10. Harms, W., Schoffel, P.J., Sroka-Perez, G., Schlegel, W., Karger, C.P.: Accuracy of a commercial optical 3D surface imaging system for realignment of patients for radiotherapy of the thorax. *Phys. Med. Biol.* 52(5), 3949–3963 (2007)
11. Tanga, J., Dieterichb, S., Cleary, K.: Respiratory Motion Tracking of Skin and Liver in Swine for CyberKnife Motion Compensation. In: SPIE Medical Imaging (2004)
12. Verellen, D., Soete, G., Linthout, N., Van Acker, S., De Roover, P., Van de Steene, J., Vinh-Hung, V., Storme, G.: Quality assurance of a system for improved target localization and patient setup that combines real time infrared tracking and stereoscopic Xray imaging. *Radiotherapy and Oncology* 67(5), 129–141 (2003)
13. Gerszten, P.C., Ozhasoglu, C., Burton, S.A., Welch, W.C., Vogel, W.J., Atkins, B.A., Kalnicki, S.: CyberKnife frameless single fraction stereotactic radiosurgery for tumors of the sacrum. *Neurosurg.* 15(2) (2003)
14. Brown, L.G.: A survey of image registration techniques. *ACM Computing Surveys* 24(4), 325–376 (1992)
15. Valet, L., Mauris, G., Bolon, P.: A Statistical Overview of recent literature in Information Fusion. *IEEE Magazine on Aeronautics and Electronics Systems* 16(3), 7–14 (2001)

Speech Signal Processing Based on Wavelets and SVM for Vocal Tract Pathology Detection*

P. Kukharchik, I. Kheidorov, E. Bovbel, and D. Ladeev

Belarusian State University, 220050 Nezalezhnasty av, 4, Minsk, Belarus

Abstract. This paper investigates the adaptation of modified wavelet-based features and support vector machines for vocal folds pathology detection. A new type of feature vector, based on continuous wavelet transform of input audio data is proposed for this task. Support vector machine was used as a classifier for testing the feature extraction procedure. The results of the experimental study are shown.

1 Introduction

Information achieved from speech analysis plays a great role in vocal tract pathology detection. In some cases such analysis is the only way to find pathologies. In medicine voice quality estimation is a very important task that caused a lot of researches in different spheres. Nowadays there are a lot of methods for direct observation and diagnostics of vocal pathologies, but they have a series of drawbacks. Human vocal tract during the sounds pronouncing is hardly observed and this is a problem for pathology detection. In addition, such examination causes discomfort to patients and influence the result reliability [1]-[2]. In this comparison the acoustical signal analysis does not have such drawbacks as pathology detection method. Except this such method has serious advantages. Firstly, acoustical signal analysis is a noncontact method, and thanks to this it lets to explore more patients in a small period of time. Secondly, it lets to detect diseases on early stages. There are done several researches in this direction based on analysis of some long vowels [3]-[4]. Last time accent in this sphere was shifted to the idea of usage of automatic speaker recognition methods for voice pathology detection [5]-[6]. The achieved accuracy is an encouraged one even for a small amount of training data. In this paper we propose the speech signals classification scheme specially developed for the vocal tract pathology detection. Base principles of this scheme are very close to those like physician analyses patient speech. As a basis for feature vector forming the continuous wavelet transform is used, and support vector machine was selected as a classifier. The main aim of this paper is to propose method for convenient continuous control of pathology evolution.

2 Methodology

Vocal pathology presence leads to changes in sounds pronunciation by a human. Depending on the pathology the changes can be more or less expressed.

* The paper is supported by ISTC grant, project B-1375.

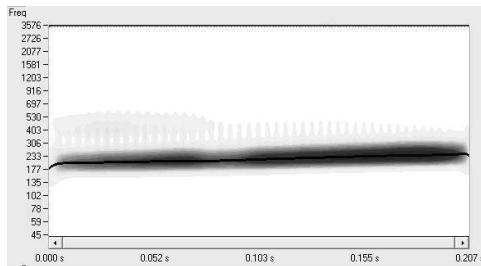


Fig. 1. Wavelet transformation of [e] sound, from the voice of speaker with normal voice

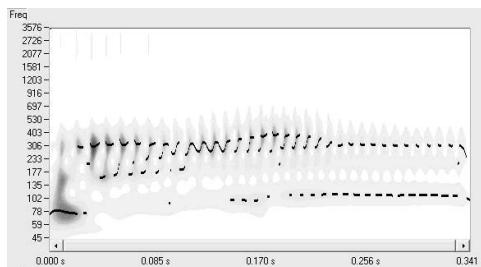


Fig. 2. Wavelet transformation of [e] sound, from the voice of speaker with polypus of vocal cord

Among sounds the most interesting are long vowels and some resonant sounds, the pathology is more evident for these sounds. On the first stage during the initial analysis the stressed vowels are to be manually selected from continuous speech and than processed by wavelet-analysis. Wavelet analysis is chosen as an optimal tool due to its effectiveness for analysis of short and non-stationary signals like phonemes. At fig.1. there is a wavelet transform of stressed sound [e] spoken by a healthy person.

If there is the pathology in a signal the picture is changed. At fig.2 there is a wavelet transform for the same vowel for patient with polypus of vocal cord. It is obvious the non-stability of fundamental frequency due to the flexibility loss by cords. It was analyzed more than 140 recordings of healthy voices and voices with pathologies, and the similar results were achieved. This fact makes us sure that wavelet transform will provide the good resolution performance for long speech fragments in order to find distortions caused by pathologies. Not any spectrum estimation method can produce the required accuracy in time-frequency domain, suitable for pathology detection.

2.1 Improved Algorithm for Wavelet Transformation

The continuous wavelet transform (CWT) of $f(t)$ can be presented as:

$$Wf(u, s) = \int_{-\infty}^{+\infty} f(t)\Psi_{u,s}(t)dt \quad (1)$$

Where wavelet Ψ function with zero mean and stretch parameter s and shift parameter u :

$$\Psi_{u,s}(t) = \frac{1}{\sqrt{s}}\Psi\left(\frac{t-u}{s}\right) \quad (2)$$

In our work we have used for CWT calculation algorithm from [7], which implement Morlet wavelet as time-frequency functions. Firstly, we used binary version of this algorithm based on powers of 2, to achieve the highest rate. The scale parameter s was changed as $s = 2^a 2^{frac{jJ}}$, where a - current octave, J - number of voices in a octave. We used $J = 8$. Secondly, the pseudo-wavelet was realized, which combines the averaging power of Fourier transform and accuracy of classical wavelet-transform. We used exponential change of base frequency and linear change of window size. This leads to the full correspondence of frequency scales of wavelet and pseudo-wavelet transform. In this case (1) transforms into:

$$W_{pseudo}f(u, s) = \int_{-\infty}^{+\infty} f(t)\rho_s(t-u)(t)dt \quad (3)$$

where $\rho_s(t)$ is a complex pseudo-wavelet with base frequency coordinated with wavelet frequency in scale s . The usage of pseudo-wavelets lets to average non-informative signal deviations during feature vector forming. In such a way we achieve higher accuracy for frequency analysis then it can be achieved using FFT.

2.2 Feature Vector

The classification scheme is shown at fig.3. The result is a time-frequency signal representation. The image of wavelet transform for each segment is the source for future feature vector extraction procedure. There are a lot of methods to construct feature vector from CWT image, but it was proposed to use the simplest one for vocal fold pathology detection task. In order to do this we use the averaging of neighbor wavelet-coefficients on time-frequency scale. The whole time-frequency range divided on sub-ranges along time and frequency scales. Then coefficients inside each mosaic element are averaged and used as feature vector parameters (fig. 4).

2.3 Support Vector Machines (SVM)

SVM is a separating classifier, simple in its structure but effective. We use SVM for the voice pathology detection and classification as an optimal classifier. Distinction in kind of SVM to commonly used classifiers as hidden Markov models (HMM), Gaussian mixture models (GMM), is that SVM directly approximates between-class borders, not modeling probability distributions of training sets. SVM classifier is defined by the elements of the training set. But not all the elements are used for the classifier creation. Usually support vector's share is not big and classifier will be thinned. Training set defines the complexity of the classifier. Classification using SVM model is a simply calculation of vector relation to the border between classes, which was built during the training procedure.

Using the SVM as classifier for the task of vocal tract pathology detection is righteous due to following reasons:

- Speech signals classification for the task of voice pathologies detection can be described as a set of two-class classifications. Classifier structure in this case is a tree, where the first class contains of the most similar in structure pathologies and second class contains all others. Then classification in every of the classes is performed. It is also to perform classification of more than two classes optimizing SVM so, that all classes are processed simultaneously [8].
- Training sequence determines complexity and accuracy of the classifier. In our experiment we use feature vectors as training elements. Bigger differences between each element of two class's vectors make easier to build classes boundaries with the SVM classifier. Space dimension is equal to the dimension of the feature vectors.
- Recognition quality is sensible to the samples topology: compact distribution of the same class samples can help the recognition task. However, wider distribution of the samples leads to the recognition difficulties. Euclidian distance cannot help solving this problem. Training sequence should be well balanced. First, number of the records of both classes should be comparable. If one class is represented with much more records than another, classifier cannot build class boundaries correctly, and misclassification rate will be high. Each record contribution in the training sequence also has to be controlled to be equal to others, and all pathologies are represented adequately.

3 Experiment

For the common case, experiment of pathology recognition task consist of:

- Database creation. Database for pathology detection and recognition must contain records of many people with different types of pathologies and without any pathology. It is better if database contains records made on different languages, so classifier effectiveness and robustness can proved.

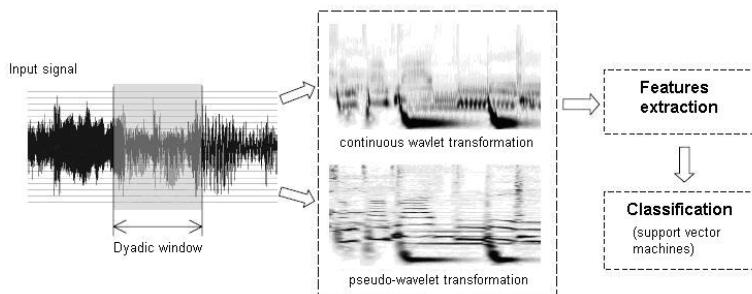


Fig. 3. Classification scheme using continuous wavelet transformation and SVM

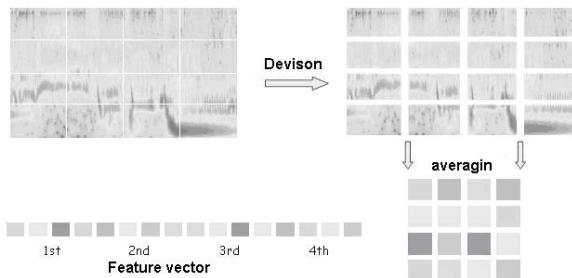


Fig. 4. Feature vector creation

- Choosing speech signal parameters for feature vector creation. Former we must specify acoustic signal type and classifier structure.
- Creation of the model for good and pathology voices using database. Former we choose learning and parameters optimization procedures.
- Model evaluation. Data is separated into two parts: learning sequence and testing sequence. Learning part we use for model creation, testing sequence we use for evaluation.
- Using real voice signals for system evaluation. It can be speech of anybody in appropriate format.

3.1 Database Description

We use database which was created in Republic Center of Hearing, Voice and Speech Pathologies (Minsk, Belarus). All records represented in audio format PCM WAVE with 44 kHz sample rate and 16 bit sample size, mono. Patients were asked to read some text during several minutes. There were no any requirements about pronunciation, clearness articulation. Patients also didn't need to pronounce long vowels. Each record was specified a diagnose made by a phoniatrist after a patient check up using special equipment. Thus was created database of around 70 hours for good voices and around 20 hours of voices with pathologies. What distinguishes this database from others (for example free available database from Massachusetts hospital lab of voice and hearing) is that our database contains patterns for natural spontaneous voice records without pre-processing. Using this database guarantees good resembling of the experiment conditions to the situation of natural voice in noisy environment. Database was created of 90 speakers: 30 speakers with the normal voices, 30 speakers with the vocal cords neps and 30 speakers with the functional pathologies. All phrases have been processed with the speech-detector and contain just numbers (from "2" to "9").

3.2 Experimental Protocol

During the experiment speech signal was divided into separate words. Each word was parameterized and represented with 8×8 and 16×4 feature vectors of

Table 1. Classification of the normal voices and voices with vocal cord neps

WORD	INPUT SIGNAL	OUTPUT SVM 8×8		OUTPUT SVM 16×4	
		correct classification	wrong classification	correct classification	wrong classification
"2"	normal (20)	16	4	19	1
	pathology(20)	17	3	20	0
"3"	normal (20)	14	6	19	1
	pathology(20)	17	3	20	0
"4"	normal (20)	19	1	19	1
	pathology(20)	17	3	20	0
"5"	normal (20)	19	1	19	1
	pathology(20)	20	0	20	0
"6"	normal (20)	19	1	19	1
	pathology(20)	20	0	20	0
"7"	normal (20)	19	1	19	1
	pathology(20)	20	0	20	0
"8"	normal (20)	19	1	19	1
	pathology(20)	20	0	20	0
"9"	normal (20)	19	1	19	1
	pathology(20)	20	0	20	0
ALL	normal (160)	144(90.0%)	16(10.0%)	152(97.5%)	8(2.5%)
	pathology(160)	151(94.3%)	9(5.7%)	160(100%)	0(0.0%)

continuous wavelet transformation: in time-frequency domain each word is divided into 8 segments along time axis and 8 along frequency axis, and averaging is performed for each of 64 2D segments. In case of 16×4 feature vector the word is divided into 16 segments along frequency axis and 4 segments along time axis. Two SVM models were trained for the classification of the records belonging to speakers with the normal voices and speakers with the pathologies: model for the classification of the normal voices and voices with the vocal cords neps, model for the classification of the normal voices and voices with the functional pathology. Testing sequence went through the classifiers and according to the output segment belonging is decided.

3.3 Experimental Results

Table 1 presents results of classification of the normal voices and voices with the vocal cord neps. Correct classification rate reached for this task using continuous wavelet transformation feature vector of size: $8 \times 8 - 92.2\%((144 + 151)/(160 + 160))$ $16 \times 4 - 97.5\%((152 + 160)/(160 + 160))$. It can be noticed from the results that vector size 16×4 is preferable for the task of pathology detection.

Table 2 presents results of classification of the normal voices and voices with functional pathology. Correct classification rate reached for this task using continuous wavelet transformation feature vectors of size: $8 \times 8 - 93.4\%((145 + 154)/(160 + 160))$ $16 \times 4 - 97.5\%((152 + 160)/(160 + 160))$

Table 2. Classification of normal voices and voices with functional pathologies

WORD	INPUT SIGNAL	OUTPUT SVM 8 × 8		OUTPUT SVM 16 × 4	
		correct classification	wrong classification	correct classification	wrong classification
"2"	normal (20)	15	5	19	1
	pathology(20)	18	2	20	0
"3"	normal (20)	16	4	19	1
	pathology(20)	18	2	20	0
"4"	normal (20)	19	1	19	1
	pathology(20)	18	2	20	0
"5"	normal (20)	19	1	19	1
	pathology(20)	20	0	20	0
"6"	normal (20)	19	1	19	1
	pathology(20)	20	0	20	0
"7"	normal (20)	19	1	19	1
	pathology(20)	20	0	20	0
"8"	normal (20)	19	1	19	1
	pathology(20)	20	0	20	0
"9"	normal (20)	19	1	19	1
	pathology(20)	20	0	20	0
ALL	normal (160)	145(90.6%)	15(9.4%)	152(97.5%)	8(2.5%)
	pathology(160)	154(96.2%)	6(3.8%)	160(100%)	0(0.0%)

Certain decreasing in classification rate takes place in case of the type of pathology: the neps of the vocal cords or the functional pathology. For the case of pathology presence detecting (normal voice or pathological voice) correct classification reaches 90%. Archived results can be considered as encouraging for reasons:

- They show that pathology information can be caught by continuous wavelet transformation and SVM classifier even though there is a few speech material is available.
- It is possible to caught not just pathology presence but also predict the type of the pathology.

4 Conclusion

This article investigates the task of pathology recognition in voice signals using wavelets and SVM. It has been shown that acoustic analysis of recorded voices is capable of making decision about pathology presence and type in the signal. Building feature vectors from wavelet transformations is a very promising approach for the task of voice pathology detection. Adjusting parameters of the classifier to the optimal levels provides acceptable precision of normal and pathology voices classification. Obtained results prove that the proposed approach is able to work in case of not sufficient amount of learning data as

well. Following work in the defined direction will be devoted to recognition rate increasing using different types of SVM classifiers and signal parameterizations.

References

1. Alonso, J.B., de Leon, J., Alonso, I., Ferrer, M.A.: Automatic Detection of Pathologies in the Voice by HOS Based PArameters. EURASIP Journal on Applied Signal Processing 4, 275–284 (2001)
2. Gavidia-Ceballos, L., Hansen, J., Kaiser, J.: A Non-Linear Based Speech Feature Analysis Method with Application to Vocal Fold Pathology Assessment. IEEE Trans. Biomedical Engineering 45(3), 300–313
3. Manfredi, C.: Adaptive Noise Energy Estimation in Pathological Speech Signals. IEEE Trans. Biomedical Engineering 47(11), 1538–1543 (2000)
4. Wallen, E.J., Hansen, J.H.: A Screening Test for Speech Pathology Assessment Using Objective Quality Measures. In: ICSLP 1996, vol. 2, pp. 776–779 (1996)
5. Fredouille, C.: Application of Automatic Speaker Recognition techniques to pathological voice assessment (dysphonia). In: Proc. of Eurospeech (2005)
6. Maguire, C.: Identification of voice pathology using automated speech analysis. In: Third International Workshop on Models and Analysis of Vocal Emission for Biomedical Applications, Florence, Italy (2003)
7. Mallat, S.: A wavelet tour of signal processing. Academic, San Diego (1998)
8. Cristianini, N., Shawe-taylor, J.: Introduction to Support Vector Machines, p. 139. Cambridge University Press, Cambridge (2001)

Pair Correlation Integral for Fractal Characterization of Three-Dimensional Histograms from Color Images

Julien Chauveau, David Rousseau, and François Chapeau-Blondeau

Laboratoire d'Ingénierie des Systèmes Automatisés (LISA),
Université d'Angers, 62 avenue Notre Dame du Lac, 49000 Angers, France

Abstract. The pair correlation integral is used to assess the intrinsic dimensionality of the three-dimensional histogram of RGB color images. For application in the bounded colorimetric cube, this correlation measure is first calibrated on color histograms of reference constructed with integer dimensionality. The measure is then applied to natural color images. The results show that their color histogram tends to display a self-similar structure with noninteger fractal dimension. Such a fractal organization in the colorimetric space can have relevance for image segmentation or classification, or other areas of color image processing.

Keywords: Color image, Color histogram, Fractal, Pair correlation integral, Feature extraction and analysis.

1 Introduction

In image processing, histograms of the pixel values are useful to many purposes [1]. For instance, they can serve for the characterization and correction of image acquisition, or for segmentation operations, or for indexing in image data bases. For gray-level or monocomponent images, the histogram is a simple one-dimensional data structure, straightforward to visualize and to handle. However, multicomponent images, including multispectral images, are becoming more and more pervasive. The histogram of a multicomponent image, accordingly, is a multidimensional data structure, and multidimensional data can exhibit complex organization [2]. For multicomponent images, the approach which is very often followed in current image processing, is to consider separately the monodimensional marginal histogram of each component. This leads to simple processing, but which may lose important information contained in the dependence between components. The full multidimensional histogram contains richer information, but it is more complicated to handle, and little is known concerning the structure of the multidimensional histogram of multicomponent images, starting with RGB color images. In the present paper, we are going to show that the histograms of natural RGB color images tend to exhibit nontrivial multiscale organization, with self-similarity across scales, and a fractal character. This seeks to contribute in the direction of better knowledge and more effective exploitation of multidimensional histograms of images.

Other fractal properties have already been reported for images, but these essentially concerned the spatial organization of images. Especially, for natural images, inherent multiscale structures and details existing in natural scenes, tend to induce fractal properties detectable in the spatial organization of the images. This is for instance conveyed by self-similar power-law evolutions observed for the spatial-frequency spectra and the associated spatial correlation functions of many natural images [34, 56, 7]. The fractal properties we are going to report here are distinct. They are not observed in the spatial organization of the images, but in their colorimetric organization, and these are new observations. The colors that are present in an image, and the frequency of their occurrences, tend also to distribute over the colorimetric space in a fractal, self-similar, way.

2 Pair Correlation Measures

We consider an RGB color image with N pixels. Each one of the three (R, G, B) components can vary among Q possible values, from 0 to $Q - 1$. In the colorimetric RGB space, each pixel, indexed by n , maps into a triplet defining a point $\mathbf{X}_n = (R, G, B)$, for $n = 1$ to N . The N points \mathbf{X}_n which distribute among the Q^3 colorimetric cells form the three-dimensional histogram we will be considering for the image. For a characterization of the histogram organization in RGB space, we evaluate the number of pairs of points which are separated by a distance $\leq r$. It is defined as

$$C_2(r) = \sum_{n=1}^N \sum_{n'>n}^N \Gamma(r - \|\mathbf{X}_n - \mathbf{X}_{n'}\|), \quad (1)$$

with the Heaviside function $\Gamma(u) = 1$ if $u \geq 0$ and $\Gamma(u) = 0$ otherwise. We can consider $C_1(r) = C_2(r)/N$ which is the average number of pairs per pixel of the image. This $C_1(r)$ reflects in the RGB space, for a typical pixel, the average number of neighbors within a distance $\leq r$ in the three-dimensional histogram. The largest possible pixel separation in the colorimetric cube $[0, Q - 1]^3$ is $r_{\max} = \sqrt{3}(Q - 1)$ and accordingly, at $r = r_{\max}$ the value of $C_2(r)$ saturates at $N(N - 1)/2$ which is the total number of pairs. Normalization by this total number of pairs yields $C_0(r) = C_2(r)/(N(N - 1)/2)$. This $C_0(r)$ is a non-decreasing function which starts close to zero at $r = 0$ and saturates at 1 at large r approaching r_{\max} . For points in space, such a function $C_0(r)$ is known as the correlation integral [89]. It is a measure which is used to characterize the organization of sets of points in space. For points uniformly scattered in the three-dimensional space, the pair correlation integral $C_0(r)$, or any of its non-normalized versions $C_1(r)$ or $C_2(r)$, would vary as r^3 . For more complex sets of points, a remarkable feature is to register a power-law variation as r^D with a possibly noninteger exponent D differing from 3. Such a behavior characterizes a nontrivial self-similar organization of the points. Self-similarity arises from invariance of the power-law under a change of scale: if the scale in r changes as $r \rightarrow \alpha r$, a corresponding change of scale in C_0 as $C_0 \rightarrow \alpha^D C_0$ makes the whole

structure look similar. The organization of the points in space, as characterized by the correlations $C_0(r)$ versus r , has no characteristic scale and is self-similar across scales. This confers to it a fractal character [10][11].

Pair correlation measures similar to $C_0(r)$ or $C_1(r)$ or $C_2(r)$, have been used to characterize fractal structures in various processes, for instance in chaotic attractors, or in porous media, or in diffusion-limited aggregates, or in percolation clusters, or in the distribution of stars and galaxies [8][10][11]. Here, we shall use the measure $C_1(r)$ to exhibit a fractal organization in three-dimensional histograms of color images.

3 Calibration of the Measure for Images

We want to use the average number of pairs $C_1(r) = C_2(r)/N$ from Eq. (1), in order to characterize the distribution of points of the histogram in the colorimetric cube $[0, Q - 1]^3$. First, for calibration purpose, it is appropriate to test the evolution of the correlation measure $C_1(r)$ on reference images with known statistical properties. This is especially useful because, when applied to images, there are no simple a priori models to theoretically predict the evolutions of $C_1(r)$. Even in the simple case, evoked above, of points uniformly scattered in the three-dimensional space, the simple prediction $C_1(r) \propto r^3$ is valid only in a homogeneous unbounded space. But since the colorimetric cube $[0, Q - 1]^3$ has well defined boundaries, when it is uniformly filled with points one may expect a deviation from $C_1(r) \propto r^3$, due to boundary effects. Accordingly, one may expect the behavior $C_1(r) \propto r^3$ to be valid at small r and to gradually disrupt as r approaches the linear size $Q - 1$ of the colorimetric cube.

For calibration of the correlation measure $C_1(r) = C_2(r)/N$ from Eq. (1), we start with the simple random image with size $N = 256 \times 256$ pixels, for which the color components R , G and B at each pixel are randomly picked in $[0, Q - 1 = 255]$ with uniform probability. A resulting realization of such a random image is shown in Fig. 1. The corresponding pair correlation measure $C_1(r)$ has been evaluated and is also presented in Fig. 1.

The log-log plot of Fig. 1 shows an evolution as $C_1(r) \propto r^3$ which is well verified, except at large and small r . At large distance r approaching $r_{\max} = \sqrt{3}(Q - 1)$, there is a saturation of the count $C_1(r)$ which tends to grow more slowly than r^3 due to the boundary effect. At small r close to $r = 1$, the count $C_1(r)$ which operates on the discrete grid of the colors in the histogram, deviates from the model $C_1(r) \propto r^3$ which assumes a continuous distribution of points. At small r , the average number of neighbors on the discrete grid differs slightly from the volume of the sphere of radius r , and produces fluctuations of the count around a straight line behavior in log-log coordinates. These fluctuations become rapidly negligible as r increases above 1, as visible in Fig. 1. Also at small r , the statistical estimation is less accurate, due to the low average density of neighbors at small r , with the random uniform model of Fig. 1. There is a total of $N = 2^{16}$ pixels in the image, for 2^{24} colorimetric cells in the histogram; this leads to a (low) uniform density of 2^{-8} point per colorimetric cell of side 1, yielding the

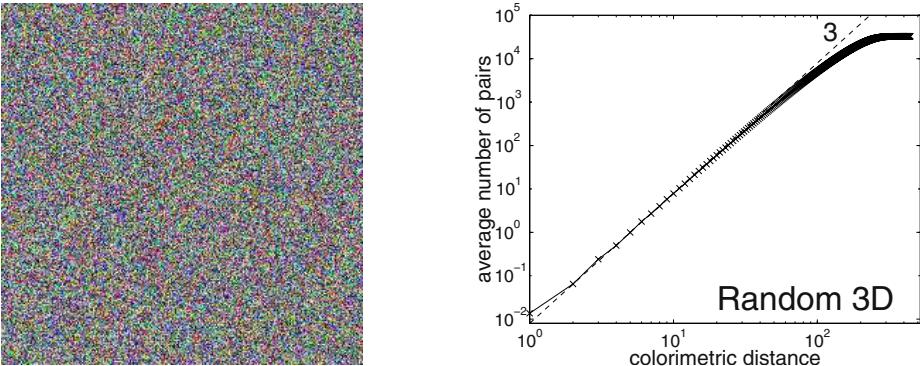


Fig. 1. Left: RGB color image with size 256×256 pixels and three random R , G and B components uniform in $[0, Q - 1 = 255]$. Right: Average number of pairs $C_1(r)$ in its three-dimensional histogram, as a function of the colorimetric distance r ; the dashed line has slope 3.

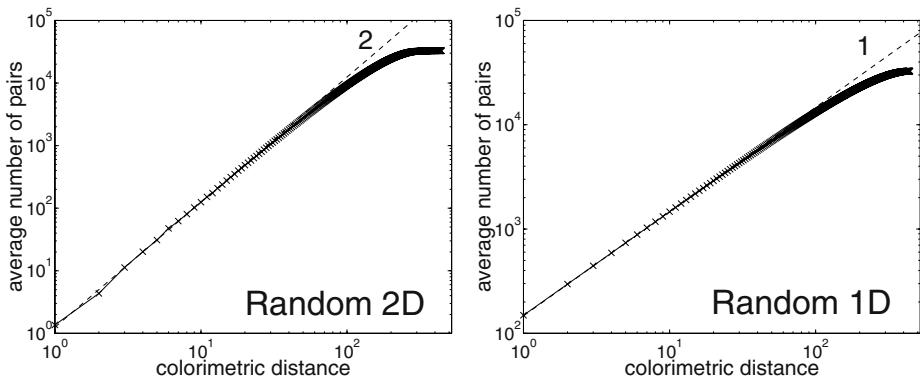


Fig. 2. Average number of pairs $C_1(r)$ in the color histogram, as a function of the colorimetric distance r , when the histogram lies in a plane (left), or on a line (right), in the colorimetric cube $[0, Q - 1 = 255]^3$. The dashed lines have slope 2 and 1.

order of magnitude of 10^{-2} for the average number of neighboring points at $r = 1$ in Fig. II Yet, within these limits at large and small r , the evolution as $C_1(r) \propto r^3$ is well identified in the results of Fig. II, and provides a consistent characterization of the uniform color histogram as a three-dimensional manifold.

We also tested the correlation measure $C_1(r)$ on another random test image whose color histogram is deliberately chosen as a two-dimensional manifold. This is achieved with a random image with size $N = 256 \times 256$ pixels, for which the components R and G at each pixel are randomly picked in $[0, Q - 1 = 255]$ with uniform probability; then the remaining component is constructed as $B = (R + G)/2$. This ensures that, in the colorimetric cube $[0, Q - 1]^3$, the histogram

of the random image lies in a plane: the principal diagonal plane with a vertex at $(0, 0, 0)$ and a vertex at $(Q - 1, Q - 1, Q - 1)$ in the colorimetric cube $[0, Q - 1]^3$. The corresponding pair correlation measure $C_1(r)$ is presented in Fig. 2(left).

The log-log plot of Fig. 2(left) shows an evolution of the pair correlation measure as $C_1(r) \propto r^2$, within the limits at large and small r . This again provides a consistent characterization of the uniform color histogram as a two-dimensional manifold here.

Next, we have tested $C_1(r)$ on a histogram constructed as a one-dimensional manifold. For the histogram, $N = 256 \times 256$ points are distributed at random with uniform density on the principal diagonal line joining vertices $(0, 0, 0)$ and $(Q - 1, Q - 1, Q - 1)$ of the colorimetric cube $[0, Q - 1 = 255]^3$. This corresponds to images with random uniform components R , G and B constrained by $R = G = B$. The log-log plot of Fig. 2(right) shows an evolution of the pair correlation measure as $C_1(r) \propto r^1$, providing, as before, a consistent characterization of the uniform color histogram as a one-dimensional manifold here.

Together, the results of Figs. 1-2 demonstrate the ability of the correlation measure $C_1(r)$ to provide a consistent identification of the dimension of the manifold formed by the color histogram in the colorimetric cube. These results, obtained with random test images, will serve to us as reference in order to interpret the behavior of $C_1(r)$ measured on natural color images.

4 Characterization of Natural Color Images

We have considered various common RGB color images, with size $N = 256 \times 256$ pixels and $Q = 256$ levels, and represented in Fig. 3. For illustration, two three-dimensional color histograms are depicted in Fig. 4.

For each of the color images of Fig. 3, we have computed the average number of pairs $C_1(r) = C_2(r)/N$ from Eq. (1), in order to characterize the distribution of points of the histogram in the colorimetric cube $[0, Q - 1]^3$. The resulting evolutions of $C_1(r)$ are reported in Fig. 5.

For each log-log plot in Fig. 5, a dashed line with slope D is indicated which provides a fit to the data. This corresponds for the pair correlation measure $C_1(r)$ to a power-law model as $C_1(r) \propto r^D$. The linear fits in Fig. 5 best apply at intermediate range of r where the practical implementation of the correlation integral is known to be best suited for dimensionality estimation [9]. In Fig. 5, noninteger values are observed for the power-law exponent D . Such power-law behavior of the pair correlation integral, with noninteger exponent D , identifies a nontrivial self-similar organization characterizing a fractal structure [9][10][11]. For the three-dimensional color histograms, this reveals an organization with no characteristic scale. The points in the histogram tend to form clusters with no characteristic size, but clusters with many sizes occurring in a self-similar way, and assessed by a noninteger fractal dimension D . In association, voids with no points may be expected to exist at all scales in the fractal histogram.

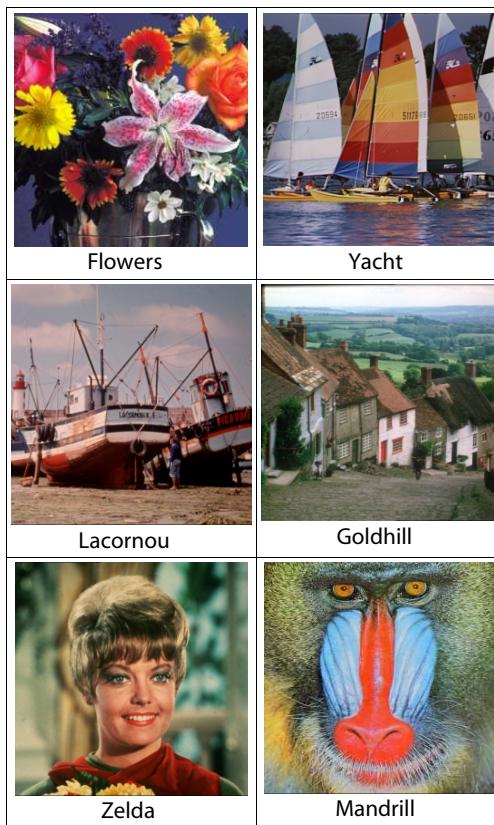


Fig. 3. Six RGB color images with size 256×256 pixels, and $Q = 256$ levels

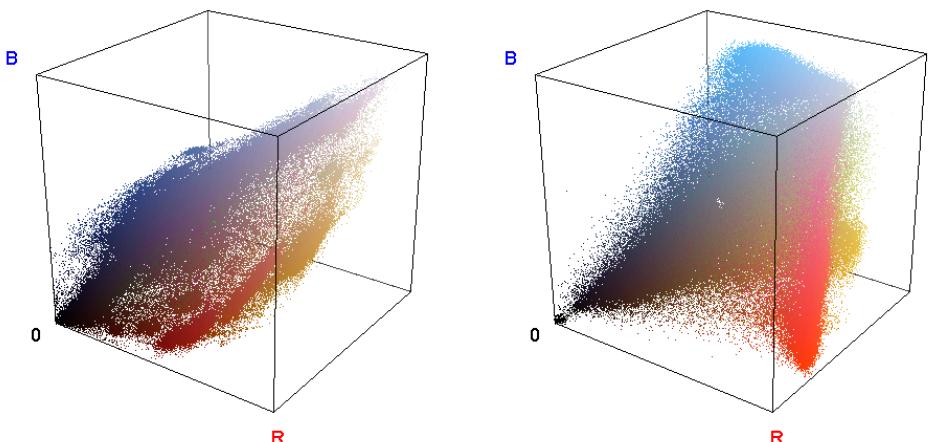


Fig. 4. Color histogram in the RGB colorimetric cube $[0, Q - 1 = 255]^3$ for image "Yacht" (left) and image "Mandrill" (right) from Fig. 3

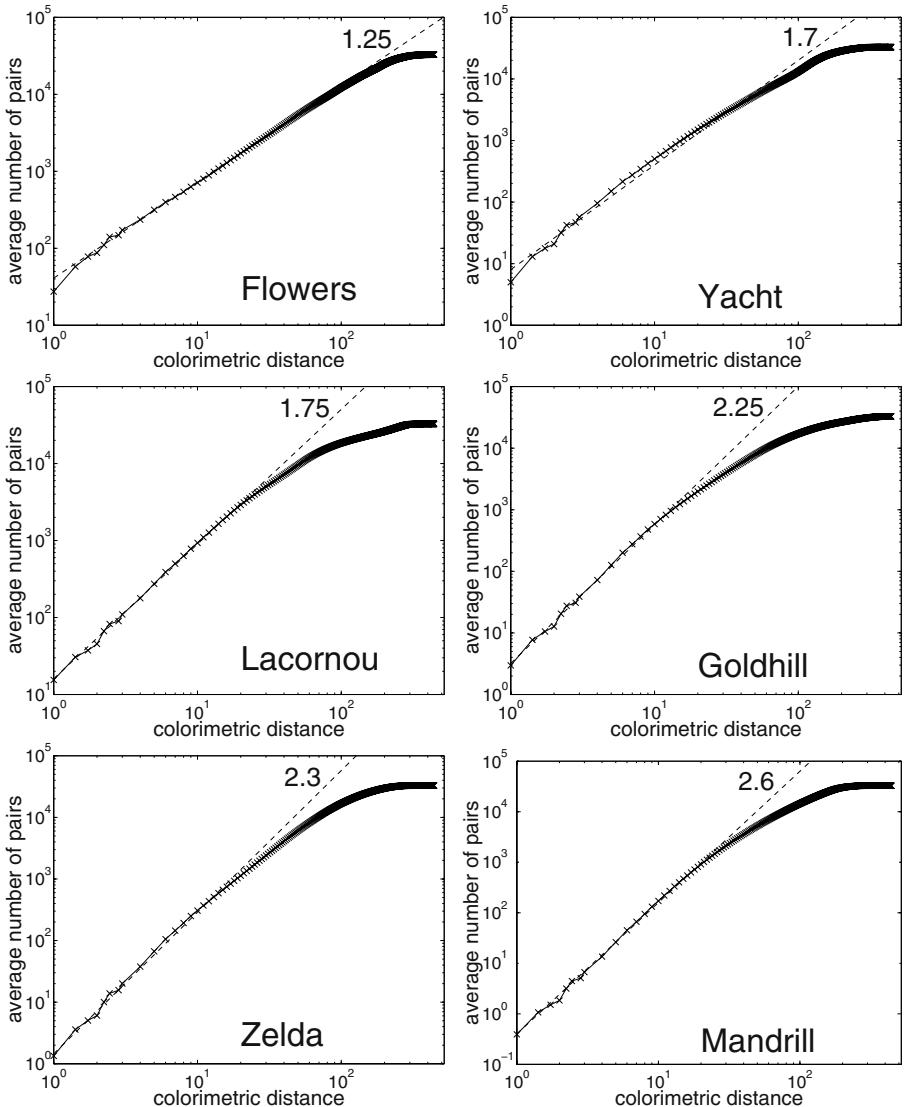


Fig. 5. Average number of pairs $C_1(r)$ in the color histogram, as a function of the colorimetric distance r , for the six RGB color images of Fig. 3. The slope is indicated of the dashed line fitted to the data.

5 Discussion

In this study, we have used the pair correlation integral estimated via Eq. (II), in order to assess the intrinsic dimensionality of the three-dimensional histogram of RGB color images. For application in the bounded colorimetric cube, the

correlation measure has been first calibrated on color histograms of reference constructed with integer dimensionality. The pair correlation integral has then been evaluated on natural color images. Our observation is that for natural images with sufficient colorimetric variability, the three-dimensional histogram tends to display a self-similar structure with a noninteger fractal dimension which is specific to a given image. The registration of possible fractal structures in the color distribution of natural images is a new feature. The preliminary results presented here open a way in this direction. Additional observations and analyses are needed, in order to confirm the colorimetric fractal properties on more extensive surveys of natural images, and in order to appreciate the conditions of existence and the possible origins of them. Other approaches or estimation methods can also be used for alternative characterizations of fractal colorimetric properties in images. We currently have complementary studies under way in this direction.

We again emphasize that the present fractal properties observed in the color distribution of natural images are different from, but possibly related to, other fractal properties previously reported for natural images in their spatial organization [3][4][5][6][7]. Briefly stated, previous results [3][4][5][6][7] dealt with fractal distribution of the pixels in space, while the present results deal with fractal distribution of the pixels in the colorimetric cube.

The registration of possible fractal structures in the colorimetric organization of natural images may have several significant consequences for color image processing. Estimating the intrinsic dimensionality of data sets, as done here for color images, is helpful to determine the possibility of new coordinate systems allowing more parsimonious representation. Also, a self-similar fractal organization with no characteristic scales, may have bearing for various processing performed with the color histograms: for instance, when implementing segmentation methods defining classes by looking for peaks in the color histogram, or when implementing subquantization of the color components or color reduction operations. In addition, the fractal dimension D may serve as a useful index for various image processing tasks, such as image characterization, classification or indexing. All these issues related to a possible fractal organization of the colors, may form interesting areas to explore for color image processing [12] and enlarge the applications of fractal geometry to imaging [13].

Acknowledgement. Julien CHAUVEAU acknowledges support from *La Communauté d'Agglomération du Choletais*, France.

References

1. Russ, J.C.: *The Image Processing Handbook*. CRC Press, Boca Raton (1995)
2. Landgrebe, D.: Hyperspectral image data analysis. *IEEE Signal Processing Magazine* 19(1), 17–28 (2002)
3. Burton, G.J., Moorhead, I.R.: Color and spatial structure in natural scenes. *Applied Optics* 26, 157–170 (1987)

4. Ruderman, D.L., Bialek, W.: Statistics of natural images: Scaling in the woods. *Physical Review Letters* 73, 814–817 (1994)
5. Ruderman, D.L.: Origins of scaling in natural images. *Vision Research* 37, 3385–3398 (1997)
6. Turiel, A., Parga, N., Ruderman, D.L., Cronin, T.W.: Multiscaling and information content of natural color images. *Physical Review E* 62, 1138–1148 (2000)
7. Hsiao, W.H., Millane, R.P.: Effects of occlusion, edges, and scaling on the power spectra of natural images. *Journal of the Optical Society of America A* 22, 1789–1797 (2005)
8. Grassberger, P., Procaccia, I.: Characterization of strange attractors. *Physical Review Letters* 50, 346–349 (1983)
9. Camstra, F.: Data dimensionality estimation methods: A survey. *Pattern Recognition* 36, 2945–2954 (2003)
10. Mandelbrot, B.B.: *The Fractal Geometry of Nature*. Freeman, San Francisco (1983)
11. Gouyet, J.F.: *Physics and Fractal Structures*. Springer, Berlin (1996)
12. Sharma, G. (ed.): *Digital Color Imaging Handbook*. CRC Press, Boca Raton (2003)
13. Turner, M.J., Andrews, P.R., Blackledge, J.M.: *Fractal Geometry in Digital Imaging*. Academic Press, New York (1998)

Accelerating Integral Histograms Using an Adaptive Approach

Thomas Müller, Claus Lenz, Simon Barner, and Alois Knoll

Technische Universität München,
Dept. of Informatics VI, Robotics and Embedded Systems,
Boltzmannstr. 3, DE-85748 Garching, Germany
`{muellenth, lenz, barner, knoll}@cs.tum.edu`

Abstract. Many approaches in computer vision require multiple retrievals of histograms for rectangular patches of an input image. In 2005 an algorithm to accelerate these retrievals was presented. The data structure utilized is called *Integral Histogram*, which was based on the well known *Integral Image*.

In this paper we propose a novel approximating method to obtain these integral histograms that outperforms the original algorithm and reduces computational cost to more than a tenth. Alongside we will show that our adaptive approach still provides reasonable accuracy – which allows dramatic performance improvements for real-time applications while still being well suited for numerous computer vision tasks.

Keywords: Computer Vision, Object Recognition, Tracking, Early Processing, Integral Histogram, Adaptive Approximation.

1 Introduction

In statistics one certainly comes along histogram computation, as a histogram is a statistical description of a set of observations. Each observation is categorized and the number of observations matching each category are summed up. If the counters on the categories are each divided by the total number of observations, the total sum of all values is 1.0 and we call the histogram *normalized*. A normalized histogram can be seen as an approximation of the probability density function from the observation data given.

If more than one aspect defines a category, observations have to be examined regarding any of the relevant aspects and thus the resulting histogram has multiple dimensions. Normalized histograms with multiple dimensions are sometimes also called *Joint Histograms* [1].

Besides visualization of statistics, one of the most relevant applications for histograms is within computer vision (CV). Here a histogram is a very important tool that can be used for various tasks like image understanding and tracking (e.g. applying *Particle Filters* [2]). For the reason of further analysis numerous different types of histograms can be computed on image data, for example color histograms [3], gradient histograms [4], color cooccurrence histograms [5], local

feature histograms [6,7] or histograms of coefficients from wavelet transforms on localized object parts [8].

In this paper we confine ourselves to the computation of color histograms, but in fact the approach presented here can be extended to any other type of histogram. When computing a color histogram the observations mentioned above refer to the pixels of the image and the categories are determined by the gray values (one-dimensional histogram) or color values (multidimensional histogram). In order to obtain an image histogram typically the number of possible values for a color is limited, for example, a three-channel color value with 8-bit color-depth (256 different values per channel) – like standard RGB color space – could be quantized into $N = 4$ bins per channel. The histogram computed for such an input image then has three dimensions, which results in a total number of $N^3 = 64$ categories.

The approach presented here proposes a new method for retrieval of *Integral Histograms* (see Section 2). Our approach is considerably faster compared to the original algorithm, because it approximates the integral histogram using an adaptive stop criterion (see Section 3). In this way, we overcome the problem of expensive initialization, a major drawback of the original algorithm, while still exploiting the benefit of extremely fast retrieval times (see Section 4).

2 The Integral Histogram

For sophisticated tasks in computer vision often histogram retrievals of rectangular regions are necessary [9]. A model-based object recognition approach could for example define a set of spatially connected histograms as a model describing the object to search. Within the recognition algorithm a magnitude of a view dozens to several thousand hypothesis (considering different rotations and scaling) have to be evaluated in order to detect an object with high accuracy. Each of the evaluations requires the retrieval of histogram(s) of a rectangular sub-image. Porikli [10] recently presented a method to speed up these retrievals dramatically. He introduced the concept of *Integral Histograms*, which is based on the *Integral Image* data structure earlier described by Viola and Jones [11].

Considering a linear sequence of observations, an integral histogram in principle specifies a bundle of histograms each summarizing observations from the beginning to a certain observation. The algorithm for one-dimensional data thus can be defined recursively by:

$$H(x^i, b) = H(x^{i-1}, b) \cup Q(f(x^i)) \quad (1)$$

For image data, which is a two-dimensional data-plane rather than an observation sequence, the computation of subsequent histograms can be specified as shown below, according to [10]:

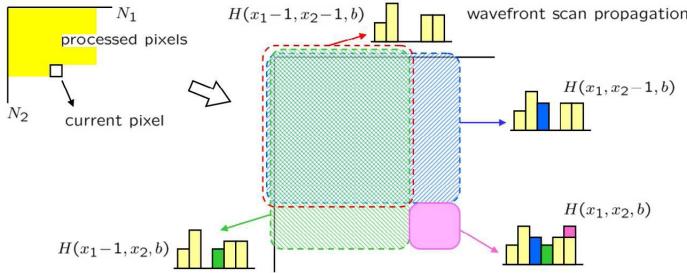


Fig. 1. Propagation of integral histogram by wavefront scan¹

$$\begin{aligned} \forall b : H(x_1, x_2, b) = & H(x_1 - 1, x_2, b) \\ & + H(x_1, x_2 - 1, b) \\ & - H(x_1 - 1, x_2 - 1, b) \\ & + Q(f(x_1, x_2)) \end{aligned} \quad (2)$$

In the above equations b denotes a bin (or category) within the histogram and $Q(f(\mathbf{x}))$ an empty histogram except for one entry that refers to the categorized value of the pixel at position \mathbf{x} . Figure 1 illustrates the flow of the algorithm.

The big advantage of the integral histogram approach is the dramatic improvement on the cost of retrieval of histograms for rectangular regions. After the first step, the computation of the integral histogram, no further image access is needed. A histogram for an arbitrary sub-image with left-upper coordinate $(x_1|x_2)$ and right-lower coordinate $(x_3|x_4)$, relevant for the evaluation of a hypothesis, can be retrieved in constant time independent from the patch size. The scheme below describes the retrieval operation from the integral data structure:

$$\begin{aligned} \forall b : H(x_1, x_2, x_3, x_4, b) = & H(x_3, x_4, b) \\ & - H(x_3 - x_1, x_4, b) \\ & - H(x_3, x_4 - x_2, b) \\ & + H(x_3 - x_1, x_4 - x_2, b) \end{aligned} \quad (3)$$

As one can see from (3), only one addition and two subtractions (in addition to array-value retrievals) are needed per bin b in order to create the new histogram. In contrast, the naive implementation needs $(x_3 - x_1) \cdot (x_4 - x_2)$ pixel queries and extra operations for computing the corresponding bin and summing up.

Obviously a big deficit in systems utilizing exhaustive histogram retrieval is, that the cost for the intersection is dependent on the size of the patch and grows proportionally to the area. Now, with the use of an integral histogram, the cost for the retrieval remains constant for arbitrary sub-images.

In a nutshell, integral histograms have great computational advantages compared to a naive approach in scenarios, where a multitude of histograms for

¹ Taken from [10]. Copyright © Mitsubishi Electric Research Laboratories, Inc., 2005

image-patches have to be retrieved, because the retrieval operation is very efficient. But still there is one big drawback which is addressed in this paper: the first step, the computation of the integral histogram, is in deed rather time and memory consuming.

3 Advanced Integral Histogram Computation

In Porikli's original work on integral histograms two methods for propagation were proposed: the wavefront scan (see Figure 1) and a string scan method. Both methods and various variations require the analysis of every single pixel of the input image, so the larger images get, the more effort is needed for computation of the integral histogram (proportional to the area).

This exact solution is not optimal by means of performance considering natural images, as in such images often uni-colored areas occur. We exploit this finding and introduce a novel method based on iterative approximation applying an adaptive stop criterion.

3.1 Adaptive Refinement

Within the iterative algorithm the input image I is equally divided into four rectangles and with every subsequent level these rectangles are divided the same way. We call this *refinement*. Intuitively recursive refinement stops at a maximum depth of

$$d_{max} = \lfloor \log_2 (\text{argmin}_I (width_I, height_I)) \rfloor. \quad (4)$$

A more sophisticated approach considering the presence of uni-colored areas defines an adaptive abort criterion. Here refinement stops, when either d_{max} is reached or the criterion is matched. The criterion $c(\mathbf{p})$ comparing the n binned values $b(p)$ of a set of pixels $\mathbf{p} = \{p_1, p_2, \dots, p_n\}$ can be defined as follows:

$$c(\mathbf{p}) = \begin{cases} \text{true} & \text{if } \forall i, j \leq n : b(p_i) = b(p_j) \\ \text{false} & \text{otherwise} \end{cases} \quad (5)$$

From (5) one can deduce that refinement is aborted, if the values of all pixels used for determination (random or equally distributed pixels inside the rectangle) are categorized into the same bin. Thus uni-colored areas are not refined any further (Figure 2).

3.2 Histogram Retrieval

As we always refine by dividing a rectangle into four sub-rectangles successively, a *Quad-Tree* seems to be a good choice for data-maintenance, because we do not loose information about the accuracy level of an approximation. Also spatial information for the histogram origins are stored efficiently. Thus from a leaf position inside the quad-tree we are able to reconstruct the position of the corresponding pixel in the original image and vice versa. Moreover, downwards from



Fig. 2. Using an adaptive stop criterion preferentially refines where strong gradients occur

the root-node we can access any histogram at a maximum of d_{max} steps, which is the maximum depth of the tree.

In order to obtain a histogram stored in a node of the quad-tree, we define a method for retrieval of a node corresponding to a position p as follows:

```

function getNode(Position p, Node root) : Node
Node n := root
int d  := 0
do
  if p = n.pos or n.noChildren then break
  else
    Node lu := n.child(left, upper)
    HVal h := p.x <= lu.pos.x ? left : right
    VVal v := p.y <= lu.pos.y ? upper : lower
    n := n.child(h, v)
  fi
  d := d + 1
while d < dmax
return n

```

One can see from the above listing, that this method always delivers the node containing the best approximation for a position if the exact position is not represented in the tree.

3.3 Recursive Approximation

On the first level of the recursive approximation, each of the histograms the integral histogram consists of is initialized with the bin value of the lower-rightmost pixel of the input image.

Within the next iterations, we recursively compute better approximations by refining and successive updating steps. Supposing we refined to the full extent of d_{max} without adaptive abortion, the recursive update can be defined according to Equation (2). But considering the use of an abortion criterion, we first have to determine the best approximation for the four sub-rectangular histograms we need for the update according to the method described in Section 3.2, before we can actually update. The positions corresponding to those are determined relative to the position of the parent histogram p^d and according to the level (the depth) d we process. For an image I we can calculate these positions (*left-upper* to *right-lower*) as follows:

$$\begin{aligned} p_{l,u}^{d+1} &= (p^d.x - 2^{-d}width_I, \quad p^d.y - 2^{-d}height_I) \\ p_{l,l}^{d+1} &= (p^d.x - 2^{-d}width_I, \quad p^d.y) \\ p_{r,u}^{d+1} &= (p^d.x, \quad p^d.y - 2^{-d}height_I) \\ p_{r,l}^{d+1} &= p^d \end{aligned} \quad (6)$$

Furthermore the histograms used for the update have to be weighted, again considering depth d and their position p :

$$w_I^d(p) = \frac{2^{2d} \cdot p.x \cdot p.y}{width_I \cdot height_I} \quad (7)$$

Utilizing Equations (6) and (7) and having retrieved the best approximating histograms $H'(p)$ as described in the last section, the update method for a bin b can be formalized as follows:

$$\begin{aligned} H(p^d, b, d) = & Q\left(f(p_{r,l}^{d+1})\right) \\ & - w(p_{l,u}^{d+1}) \cdot H'\left(p_{l,u}^{d+1}, b\right) \\ & + w(p_{l,l}^{d+1}) \cdot H'\left(p_{l,l}^{d+1}, b\right) \\ & + w(p_{r,u}^{d+1}) \cdot H'\left(p_{r,u}^{d+1}, b\right) \end{aligned} \quad (8)$$

Note that since $Q(f(p_{r,l}^{d+1})) = Q(f(p^d))$ refers to a single entry, its weight has to be set to 1.0 accordingly. As a final step, to keep all contributions in balance for the updates in $d - 1$, the resulting histogram has to be normalized according to the weights from (8):

$$\bar{H}(p^d, b, d) = \frac{H(p^d, b, d)}{1 - w(p_{l,u}^{d+1}) + w(p_{l,l}^{d+1}) + w(p_{r,u}^{d+1})} \quad (9)$$

4 Experimental Results and Conclusion

Figure 3 depicts mean retrieval times for a single histogram. In case of the integral histograms the values shown are computed with respect to the relatively

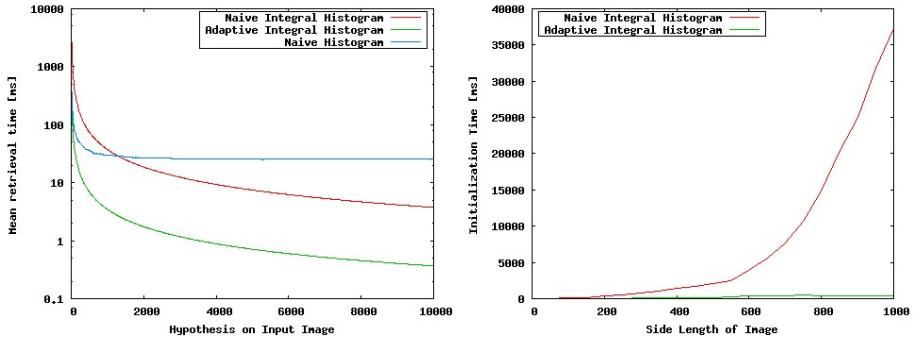


Fig. 3. Left: Mean retrieval times (note the logarithmic scale), Right: Initialization times subject to side length of the input image

expensive initialization of the integral histogram (right graph). Note that the left graph is drawn in logarithmic scale, which clarifies the enormous speed-up applying the presented algorithm. As one can see, the more hypothesis are evaluated, the less the initialization plays a role.

An important finding is, that applying the adaptive approach results in a computational benefit with a mere magnitude of about 50 hypothesis on a ~ 2.3 M pixel input image. The naive integral approach [10] reaches this break even point with above 1800 hypothesis. Mean values for retrieval of a *single* random histogram in an integral histogram structure fluctuate between 0.01 to 0.03 ms, while retrieval using a traditional approach is highly dependent on the size of the patch and takes a mean of 24.9 ms in our evaluation scenario.

Utilizing our approach the problem of expensive creation of the integral histogram can be overcome. In fact mostly initialization can be executed in around 150 ms and for smaller images $\sim 250^2$ px, considering 8-bit color depth and 8 bins per channel, the approach easily reaches the real-time limit of 25 Hz. The presented algorithm on the one hand benefits from very short retrieval times

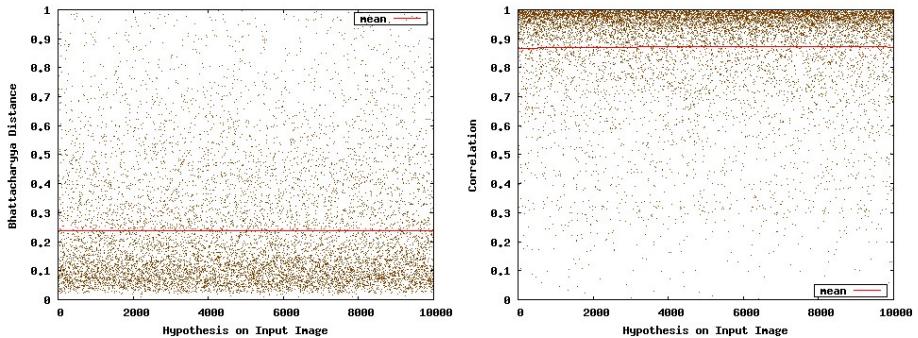


Fig. 4. Accuracy using the adaptive integral histogram approach compared to exact solutions (left: Bhattacharyya Distance, right: Correlation) on random patches

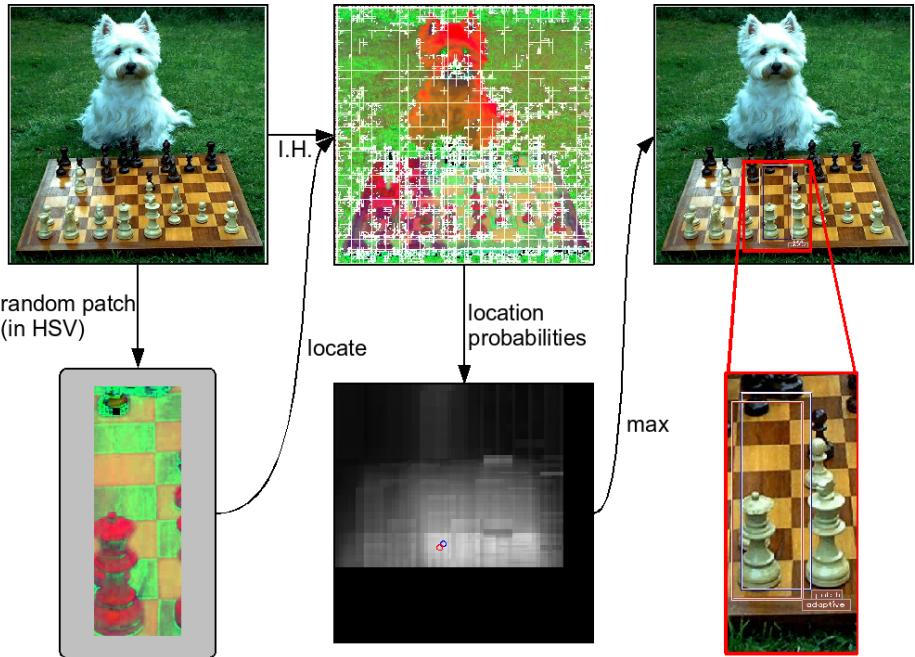


Fig. 5. Demo application with exhaustive search of a random patch

provided by the integral histogram data structure and on the other reduces the cost of structure initialization dramatically (at 2.3 M pixel by 26.27) compared to [10]. We can thus reach a speed-up proportional to the number of hypothesis compared to a conventional approach. This speed-up grew up to ~ 250 when retrieving 10^4 random histograms in our experiments.

A deficit of approximating solutions in real world applications always is the cost of achieving a desired accuracy. Concerning this topic, Figure 4 shows two common measures, the Bhattacharyya distance [12] and the correlation. Values in the figure have been compared to the exact solution obtained with the conventional approach. It becomes clear, that the solutions are well suited for real-world applications like particle filters. Moreover, due to extremely fast retrieval times, one can easily compensate for occasional outliers by simply evaluating a few more histograms.

Figure 5 shows a possible application of our approach. In this test, the adaptive integral histogram approach was used to locate a random patch in an image. The demo application uses exhaustive search and so evaluates every possible region (considering size) of the original image in order to detect the patch. The figure also shows a map of saliency - brighter areas refer to higher , darker regions to lower similarity. On the left, the extracted integral histogram using hue and saturation of the HSV color space is shown.

In a nutshell the proposed approach already outperforms the conventional approach with ~ 50 hypothesis to evaluate on an input image, while the original algorithm [10] does not until ~ 1800 on natural color images with 8 bit color-depth – 30 times faster. Yet there is some potential for improvements. Accuracy can be tuned through adjustments on the adaptive stop criterion and the number of bins. Performance on the other hand can be optimized by limiting d_{max} to a lower depth, resulting in further acceleration up to a factor 10 in experiments – a factor 300 (!) in total. Finally, due to the tree data structure, the algorithm can easily be parallelized in order to gain even more on the performance side.

Acknowledgement

This work was partly supported by the DFG excellence initiative research cluster *Cognition for Technical Systems (CoTeSys)* and the EU integrating project of Framework Program 6, *Joint-Action Science and Technology (JAST)* (FP6-003747-IP).

References

1. Pass, G., Zabih, R.: Comparing images using joint histograms. *Multimedia Systems* 7(3) (1999)
2. Arulampalam, S., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing* 50(2), 174–188 (2002)
3. Swain, M.J., Ballard, D.H.: Color indexing. *International Journal of Computer Vision* 7(1), 11–32 (1991)
4. Schiele, B., Crowley, J.L.: Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision* 36(1), 31–50 (2000)
5. Chang, P., Krumm, J.: Object recognition with color cooccurrence histogram. In: *IEEE CVPR* (1999)
6. Hetzel, G., Leibe, B., Levi, P., Schiele, B.: 3d object recognition from range images using local feature histograms. In: *IEEE CVPR*, vol. 2 II, pp. 394–399 (2001)
7. Laptev, I.: Improvements of object detection using boosted histograms. In: *BMVC 2006*, vol. III, p. 949 (2006)
8. Schneiderman, H., Kanade, T.: A statistical model for 3d object detection applied to faces and cars. In: *IEEE CVPR* (2000)
9. Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments-based tracking using the integral histogram. In: *IEEE CVPR*, vol. I, pp. 798–805 (2006)
10. Porikli, F.M.: Integral histogram: A fast way to extract histograms in cartesian spaces. In: *IEEE CVPR*, vol. I, pp. 829–836 (2005)
11. Viola, P., Jones, M.: Robust real-time object detection. In: *IEEE ICCV Workshop on Statistical and Computational Theories of Vision* (2001)
12. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.* 35, 99–109 (1943)

Volume Estimation of Small Particles Using Three-Dimensional Reconstruction from Multiple Views

Eduardo Castillo-Castaneda¹ and Christelle Turchiuli²

¹ Instituto Politécnico Nacional, CICATA,
Cerro Blanco 141, Colinas del Cimatario, 76090,
Querétaro, Mexico
ecastillo@ipn.mx

² Agroparistech, UMR GenIAI 1145, 1 avenue des Olympiades,
91744 Massy, France
christelle.turchiuli@agroparistech.fr

Abstract. This work presents a methodology and experimental results about the three dimensional reconstruction of small objects from an image sequence. The three dimensional model obtained from this methodology allows the volume estimation of particles less than 1 mm. More precisely, the particles studied were generated by an agglomeration process from milk powder. To validate this methodology, a mechanical device with six degrees of freedom was built; it allows the montage of the image acquisition system as well as the turning system that rotates the particles. The algorithm for image processing and volume estimation was implemented using a high level language.

Keywords: Volume estimation, Three-dimensional reconstruction, Multiple views.

1 Introduction

A large quantity of food products are produced or used in the form of agglomerates, appealing for its ability to "solubility." This is a mechanical property that seeks to improve the product characteristics by changing parameters of the manufacturing process. The process of the agglomeration is to combine individual particles (a few hundred microns in size) to alter the shape of agglomerate and increase the porosity, then improving the solubility and the ability of powder rehydration, facilitating the penetration of the liquid inside the agglomerate.

The methods so far used to characterize the shape of the agglomerates are based on visualization through microscopes connected to computers to obtain digital images of the particles agglomerated [1]. A highly accurate technique is the use of confocal microscope [2] which is an optical imaging technique used to increase micrograph contrast and / or to reconstruct three-dimensional images by using a spatial pinhole to eliminate out-of-focus light or flare in specimens that are thicker than the focal plane.

Measuring open porosity by mercury porosimetry, which requires the application of high pressure, it is not suitable for the characteristics of agglomerates normally fragile

that can break during measurement. The open porosity ϵ_{agg} can be also obtained from the volume envelope V_{env} , the mass m , and the density ρ of agglomerate:

$$\epsilon_{agg} = 1 - \frac{m}{\rho \cdot V_{env}} . \quad (1)$$

The volume calculation of irregularly shaped objects is usually solved by using the principle of Archimedes. However, the conventional methods fail when it comes to porous or fragile objects that cannot be subjected to external forces without altering their mechanical properties. In this case, methods of measuring volume without contact should be evaluated. A natural choice is the use of a computer-vision system that estimates the volume from a three-dimensional reconstruction of the object.

The work presented in [3] provides a rough measurement of the volume and the shape of agglomerates. The particle is placed in a turning table and the camera performs three shots at 0° , 120° and 240° . Using orthogonal projections, an average projected area of the particle is computed, then the volume of the sphere corresponding to that average area. The accuracy of this method is low because it assumes that the particles are spherical, which is not usually the case.

There are different approaches for three-dimensional reconstruction. In the most common case, it comes to getting the depth of a scene from two or more bi-dimensional images by matching techniques

The paper in [4] investigates a practical heuristics method for reconstruction of structured scenes from two uncalibrated images. This technique is used solely for the purpose of visualization a three-dimensional surface but not for purposes of calculating the volume since the information input (two images) is quite poor.

Some algorithms have been developed for modeling complex shaped objects using multiple views from image sequences [5], [6], [7]. Those works are mainly devoted to generate textured 3D models from real objects for applications such as virtual reality, computer animation and TV broadcasting. However, none of them have been particularly designed for volume measuring of small particles as the agglomerates concerned in this work.

This work presents an approach to calculate the open porosity; it consists on volume envelope estimation through three-dimensional reconstruction from multiples views of an agglomerate.

2 Methodology

The 3D reconstruction method presented here is based on the projection of multiple views of the object's silhouette to an initial volume cube that surely contains the object. The agglomerate turns on a rotary positioning device while a calibrated camera acquires an image sequence at fixed angular increments, as shown in Fig. 1. The particle turns around z-axis while N images are acquired at an angle θ_i , $i = 1, \dots, N$. The angular increment α between each view is: $\alpha = \theta_{i+1} - \theta_i$.

The images are saved in JPG format. A backlight illumination provides an easy way to extract the silhouettes from images by conventional thresholding.

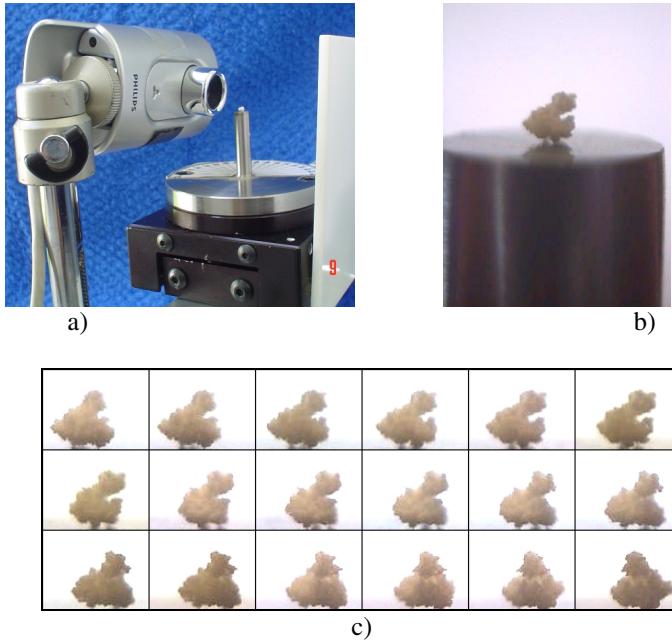


Fig. 1. a) Rotary positioning device and usb camera, b) one agglomerate zoomed view, c) multiple views at fixed angular increments (10 degrees)

The silhouette of the i image is described by a set of two-dimensional coordinates where $S_i = (ys_{i,j}, zs_{i,j})$ is the number of black pixels that belong to the silhouette. The relationship that describes the three-dimensional coordinates of the points that belong to the projection $P_i = (xp_{i,j}, yp_{i,j}, zp_{i,j})$ of the silhouette i is (considering a rotation of θ_i):

$$\begin{bmatrix} xp_{i,j} \\ yp_{i,j} \\ zp_{i,j} \end{bmatrix} = \begin{bmatrix} xs_{i,j} \\ ys_{i,j} \\ zs_{i,j} \end{bmatrix} \begin{bmatrix} \cos \theta_i & -\sin \theta_i & 0 \\ \sin \theta_i & \cos \theta_i & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (2)$$

The coordinates of the points belonging to the volume $V_i = (xv_{i,j,k}, yv_{i,j,k}, zv_{i,j,k})$ are generated by M translations T_k , $k = 1, \dots, M$, that are normal to the plane of the projection P_i , it means:

$$\begin{bmatrix} xv_{i,j,k} \\ yv_{i,j,k} \\ zv_{i,j,k} \end{bmatrix} = \begin{bmatrix} xp_{i,j} \\ ys_{i,j} \\ zs_{i,j} \end{bmatrix} + \begin{bmatrix} k\lambda \sin(\theta_i + \pi/4) \\ k\lambda \cos(\theta_i + \pi/4) \\ 0 \end{bmatrix}. \quad (3)$$

Where λ is a constant step equal to 10 to ensure enough density in the generated volume. Thus, the total volume is given by the intersection of N partial volumes V_i :

$$V_T = \bigcap_{i=1,\dots,N} V_i . \quad (4)$$

Since the silhouette projection is a pixel-based algorithm, the volume is given simply by counting the number of voxels belonging to the intersection. For a three dimensional visualization of the reconstructed agglomerates, the volume is presented as a cloud of points in the Cartesian space. A triangular mesh using the Delaunay triangulation that is a set of lines connecting each point to its natural neighbors approximates the surface of the volume model.

As an example, the Fig. 2a shows a silhouette of the Eiffel Tower. The volume generated for the projection at 0 degrees is shown in Fig. 2b. Two volumes generated at 0 and 90 degrees are shown in Fig. 2c. Finally, the three-dimensional reconstruction is shown in Fig. 2d. In this example the number of black pixels in each silhouette is 6975; the total volume is 284628 voxels.

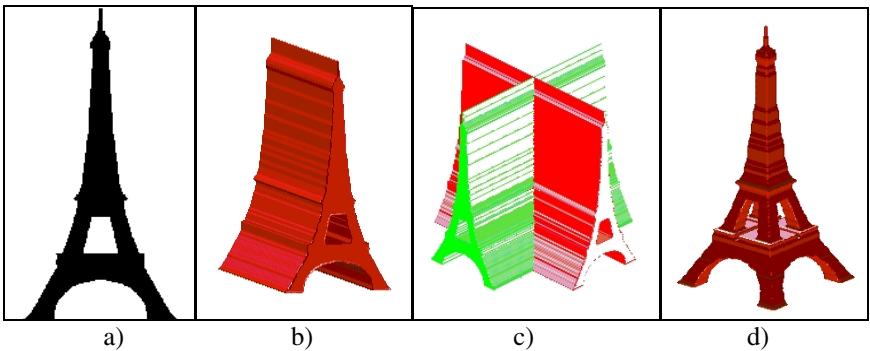


Fig. 2. a) Silhouette b) Projection and its translation, c) Two projections, d) Three-dimensional reconstruction

3 Numerical Validation

The technique of volumes intersection was implemented in c++ language. That implementation was validated using the image of a circle of 100 pixels in diameter supposed to be the image of a sphere of 523600 voxels volume. The advantage of such object is that image remains unchanged for any rotation angle of the object. Thus, it was possible to validate the accuracy of the method depending on the number of images acquired. Some of the three-dimensional images obtained are presented in Fig. 3.

Actually, there is very little difference in the quality of reconstruction for more than 20 images. Table 1 presents the estimated volumes depending on the number of images acquired, including the error of the estimation considering the actual volume of a sphere diameter of 100 pixels (actual volume equal to 523600 voxels).

The table shows that the percentage of error for more than 20 images is less than 1%. The error for 180 images (views with angular increments of 1 degree) is not very much different for 60 images (views with angular increments of 3 degrees).

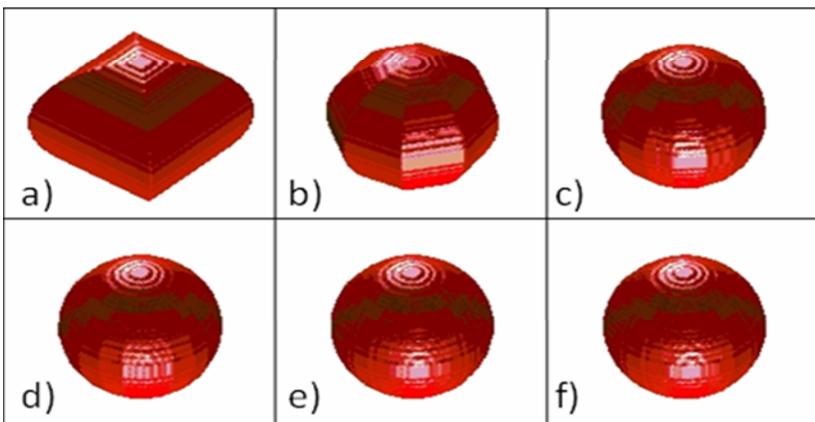


Fig. 3. Three-dimensional reconstruction with different number of images, a) 2, b) 4, c) 8, d) 10, e) 20, f) 60

Table 1. Estimated volume and error depending on the number of images

No. of images	Computed volume (voxels)	Error (voxels)	Error (%)
2	667,368	143,768	27.46
4	554,892	31,292	5.98
6	538,976	15,376	2.94
8	533,486	9,886	1.89
10	530,784	7,184	1.37
20	527,404	3,804	0.73
30	526,932	3,332	0.64
60	526,056	2,456	0.47
90	525,852	2,252	0.43
180	525,740	2,140	0.41

4 Experimental Validation and Results

To validate the performances of the angular positioning device and extraction of silhouette from actual images, one used a plastic soldier 2.5 cm in height. In this case, 18 images were used, that means angular increments of 10 degrees. The size of the images was 102x63 pixels. A backlight illumination provides good contrast between object and background that simplifies the silhouettes extraction from images. Fig. 4 presents the 18 views and the respective silhouettes extracted.

Fig. 5 presents the three-dimensional reconstruction obtained; Fig. 5a shows that there is a minimal loss of object features. A volume of 128879 voxels was obtained, whereas 1mm in the actual image is equivalent to 6.7 pixels, the volume estimated is

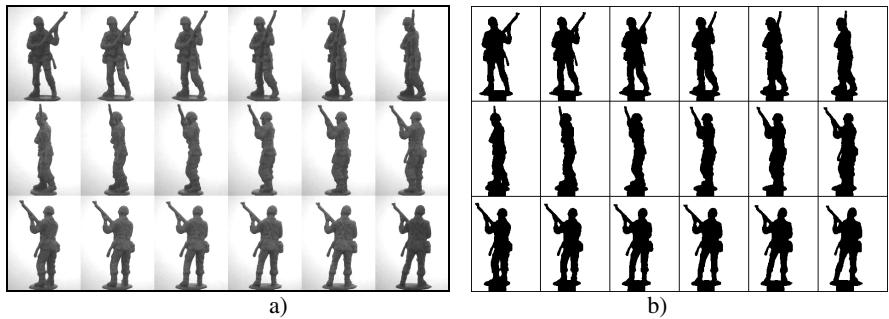


Fig. 4. a) Views of the object, b) Extracted silhouettes

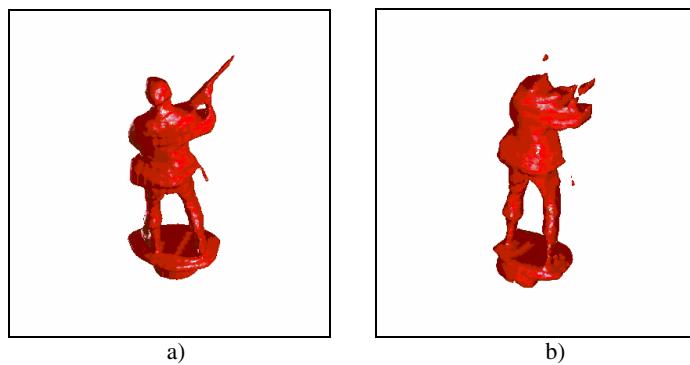


Fig. 5. a) Three-dimensional reconstruction, c) faulty reconstruction by an error in the positioning system

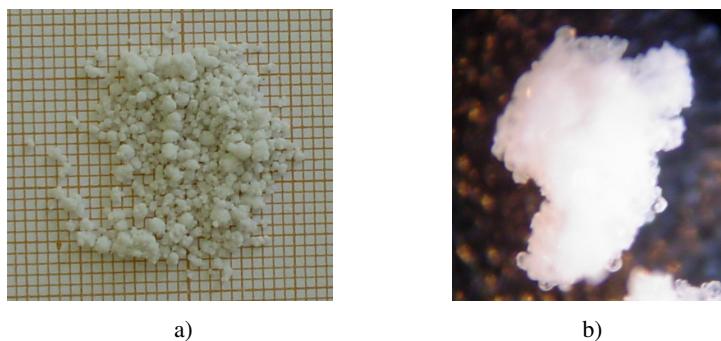


Fig. 6. a) Agglomerate sample on graph paper, b) Zoomed view of one agglomerate

428 mm³. Using the principle of Archimedes, the actual volume of the object was 410mm³. In Fig. 5b there is a faulty reconstruction obtained by an error in the orientation angle of the positioning system, causing the loss of some features of the object.

The methodology was tested in milk agglomerated particles, the size of the samples and its structure can be seen in Fig. 6.

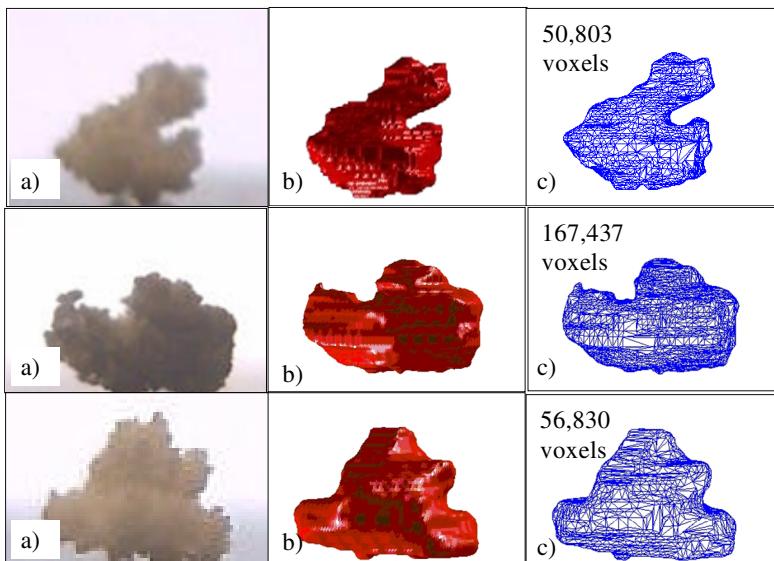


Fig. 7. Three different agglomerates: a) Actual object, b) 3D reconstruction, c) triangular mesh and volume

The Fig. 7 presents some agglomerates, their three-dimensional reconstructed models, and the estimated volume.

5 Conclusions

This paper presents an experimental validation of a technique for three-dimensional reconstruction of small particles from processing a sequence of images. The reconstruction allows the calculation of the volume of the particles.

The object digitalization generates loss of accuracy of the method, however, as demonstrated in the case of the reconstruction of the sphere, the error generated is less than 1% for a number of images around 20; the foregoing allowed to consider only 18 images, that is views with an angular increment of 10 degrees.

The calculation of each projected image on the candidate volume requires a very precise knowledge of the rotation axis position. To find precisely that position, two views of the object were used (0 and 180 degrees). The composition of both views should form a symmetrical image; the position of the axis of symmetry corresponds to the position of the rotation axis of the positioning system.

The size of the images is considered important in the final resolution of this method. The total amount of memory available on the PC where the experiments were conducted, limited the image size to 250x250 pixels.

References

1. Turchiuli, C., Eloualid, Z., El-Mansouri, N., Dumoulin, E.: Fluidised bed agglomeration: Agglomerates shape and end-user properties, Internal report ENSIA (2005)
2. Pawley, J.B.: Handbook of Biological Confocal Microscopy, 3rd edn. Springer, Berlin (2006)
3. Hogekamp, S., Pohl, M.: Porosity measurement of fragile agglomerates, Powder Technology, pp. 385–392 (2003)
4. Wang, G., Tsui, H.T., Hu, Z.: Reconstruction of structured scenes from two uncalibrated images. Pattern Recognition Letters 26(2), 207–220 (2005)
5. Fang, Y.H., Chou, H.L., Chen, Z.: 3D shape recovery of complex objects from multiple silhouette images. Pattern Recognition Letters 24(9-10), 1279–1293 (2003)
6. Fitzgibbon, A.W., Cross, G., Zisserman, A.: Automatic 3D Model Construction for Turn Table Sequences. In: Proceedings of the European Workshop on 3D Structure from Multiple Images of Large-Scale Environments, pp. 155–170 (1998)
7. Niem, W., Buschmann, R.: Automatic Modelling of 3D Natural Objects from Multiple Views. In: European Workshop on Combined Real and Synthetic Image Processing for Broadcast and Video Production, Hamburg, Germany (1994)

A Multivariate Hit-or-Miss Transform for Conjoint Spatial and Spectral Template Matching

Jonathan Weber and Sébastien Lefèvre

LSIIT, CNRS / University Louis Pasteur - Strasbourg I
Parc d'Innovation, Bd Brant, BP 10413
67412 Illkirch Cedex, France
{weber,lefevre}@lsit.u-strasbg.fr

Abstract. The Hit-or-Miss transform is a well-known morphological operator for template matching in binary and grey-level images. However it cannot be used straightforward in multivalued images (such as colour or multispectral images) since Mathematical Morphology needs an ordering relation which is not trivial on multivalued spaces. Moreover, existing definitions of the Hit-Or-Miss Transform in grey-level use only spatial templates (or structuring elements) which could be insufficient for some feature extraction problems. In this paper, we propose a multivariate Hit-or-Miss Transform operator which combines spatial and spectral patterns to perform template matching. We illustrate its relevance with an application in the remote sensing field, the extraction of coastline from very high (spatial) resolution images.

Keywords: Mathematical Morphology, Hit-or-Miss Transform, multivalued image, template matching, remote sensing.

1 Introduction

Nowadays, image sensors produce multivalued images in a wide set of applications. So the need for reliable operators adapted to multivalued data become more and more critical, in order to deal with various problems such as object detection, segmentation, classification, filtering, etc. Among these problems, object detection and recognition are of primary importance in many domains such as remote sensing, colour video processing, astronomy, and can be solved using template matching. An efficient template matching operator will help to automatically process huge collections of images. Mathematical Morphology (MM) offers several tools for image processing, including a template matching operator called the *Hit-or-Miss Transform* (HMT). This operator is effective on binary and grey-level images but has not been defined on multivalued images (such as colour or multispectral images) yet. So the contribution of this paper is to propose an adequate solution for morphological template matching in multivalued images, by defining an original multivariate hit-or-miss operator.

In this paper, we first recall main existing definitions of the Hit-Or-Miss Transform for monovalued images and discuss their possible extension to multivalued images. We believe that none of these existing approaches seems to be reliable to be extended to multivalued images in case of problems where both spatial and spectral knowledge have to be taken into account. So we introduce a multivariate Hit-or-Miss Transform relying on a set of structuring elements combining spatial and spectral descriptions. We finally illustrate the relevance of the proposed operator with the problem of coastline extraction in remote sensing.

Let us introduce notations used in this article. $D \subset \mathbb{N}^2$ is the spatial domain of images, a grey-level image is represented by a function $f : D \rightarrow \mathbb{N}$, while a multivalued image is noted $F : D \rightarrow \mathbb{N}^d$, F_b being a multivalued image band (i.e. a grey-level image f), f^- and f^+ are respectively the lowest and highest bounds of the pixel value range in f . Thus $f(p)$ represents a scalar value in case of grey-level image while $F(p)$ denotes a vectorial value in case of multivalued image. Reflection is denoted \check{f} and defined by $\check{f}(p) = f(-p)$. Duality is denoted f^* and defined by $f^* = -\check{f}$. Support of a function is defined by $\text{supp}(f) = \{p \in D \mid f(p) > -\infty\}$.

2 Monovalued Hit-or-Miss Transform

In this section, we briefly recall basic foundations of MM before reviewing existing definitions of the Hit-or-Miss Transform. Finally we explain how these definitions can be applied to multivalued images.

2.1 Foundations of Mathematical Morphology

Mathematical morphology is a theoretical framework introduced 40 years ago by G. Matheron and J. Serra [1] to compute quantitative description of geometrical structures through a spatial-based analysis. It is based on two operators called *erosion* (ε) and *dilation* (δ) respectively defined as :

$$\varepsilon_g(f)(p) = \inf_{y \in \text{supp}(g)} \{f(p + y) - g(y)\} \quad (1)$$

$$\delta_g(f)(p) = \sup_{y \in \text{supp}(g)} \{f(p - y) + g(y)\} \quad (2)$$

A pattern g called Structuring Element (SE) is involved in most of the morphological operators. We consider here the general case of functional SE, where $g : D \rightarrow \mathbb{N}$. Let us state that flat structuring elements which are used most often in MM can be expressed as particular cases of functional SEs where values of all SE's points are set to the additive identity of the value space (most often zero).

These two operators can be combined to build most of other MM operators, such as *opening* and *closing*. MM also offers a template matching operator called *Hit-or-Miss Transform* which has been defined for binary [1] and grey-level images [2].

2.2 Grey-Level Hit-or-Miss Transform

The Hit-or-Miss Transform has been first defined for binary images. In such images, this operator [1] is quite trivial and uses two disjoint structuring elements: the first has to match the foreground while the second has to match the background. Both matches are necessary in order the operator to give a positive matching response. Extension of the HMT operator to grey-level images led to several definitions which have been reviewed and unified by Naegel *et al.* [2]. The two main approaches are:

Soille :

$$HMT_{g,h}^{Soille}(f)(p) = \max\{\varepsilon_g(f)(p) - \delta_h(f)(p), 0\} \quad (3)$$

with g, h flat structuring elements.

Ronse :

$$HMT_{g,h}^{Ronse}(f)(p) = \begin{cases} \varepsilon_g(f)(p) & \text{if } \varepsilon_g(f)(p) \geq \delta_h(f)(p) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

with g, h functional structuring elements.

Ronse and Soille approaches are both composed of two steps for each pixel: first a *fitting* to check if the given pixel matches the pattern defined by the SE, and second a *valuation* to give to the fitted pixel a value in the resulting image. Fitting of these methods ensures the translation invariance property. Thus they focus on the difference between the results of the erosion and dilation. This is particularly relevant in the general case, but as soon as some template knowledge on spectral minimal or maximal values is available, the interest of these methods decreases. Moreover, matching non-uniform features composed of several parts with different spectral or intensity behaviour (e.g. plane with white fuselage, black wings parked on a grey airway) will be impossible because of the fixed number of SEs (i.e. 2).

2.3 Extension to Multivalued Images

In order to extend Soille and Ronse definitions to multivalued images, lattice theory has to be considered to formulate MM operators, thus requiring an ordering relation between pixels values. Contrary to grey-level images where pixel scalar values can be easily ordered using the single natural order, the case of multivalued is complex since various possible vectoring ordering schemes are available, none of them being widely accepted. The reader will find more details on vectorial ordering for multivariate MM in the recent survey from Aptoula and Lefèvre [3]. Two main options can be followed:

Marginal strategy : it consists in processing each band separately and independently from the others. Despite its ease of use, its disadvantages are the loss of correlated information between bands (thus the spectral signature of the pixels) and the possible lack of vector preservation (values which were not in the

input image may appear in the resulting image, thus resulting in new spectral signatures).

Vectorial strategy : it uses a vectorial ordering, such as one of the following: *conditional orderings* which give priority to some particular bands, *reduced orderings* which reduce vectors to scalar values and then order these values or *partial orderings* which cluster vectors into groups of equivalence with an ordering relation between groups. This strategy preserves vectors and use the correlated information between bands, but its main default is the need for the chosen vectorial ordering to be meaningful for the problem under consideration.

These two strategies can be used to extend grey-level HMT definitions to multivalued images. However, in the frequent case where template matching should be performed using spectral properties (thus needing to keep the correlated information) but where no meaningful vectorial ordering scheme seems adapted to the problem to be solved, both strategies may be irrelevant. Another way to perform morphological template matching in multivalued images is thus necessary.

3 Multivariate Hit-or-Miss Transform

We introduce here a new definition of the Hit-or-Miss Transform for multivalued images, thus resulting in a *Multivariate Hit-or-Miss Transform* (MHMT). We also see how this operator can be used to deal with grey-level images and point out its advantages over other existing HMTs.

3.1 Definition

The proposed MHMT is not based on marginal or vectorial strategies explained in section 2.3. Indeed a new strategy is necessary to avoid the limits of these two strategies and to correctly take into account spectral properties without selecting a (possibly inadequate) vectorial ordering. Each band is processed independently like in the marginal strategy, but with a particular SE (while the same SE is used for all bands in marginal strategy). Here each SE is dedicated to a given band (multiple SEs can be related to the same band), thus the behaviour of the HMT is specific to each band. Moreover, the HMT will return a positive result if all the SE have been correctly matched. The spectral properties are considered through the merging of the results obtained for each SE. This strategy does also not need a vectorial ordering which is particularly relevant when no vectorial ordering seems to be meaningful for the problem to be solved. Compared to the two other strategies, here the result image is not multivalued but will contain scalar values, which can be easily processed by the end-user (e.g. display, thresholding, etc).

Unlike the grey-level HMT operators presented in section 2.2, the MHMT can involve more than two SEs. It allows complex templates with very different parts (e.g. an object with red, blue, green and yellow parts) to be matched. To facilitate the understanding of the proposed HMT, we consider here flat SEs associated with threshold values instead of functional SEs. However, our transform can be directly expressed with functional SEs. Since we involve threshold values,

not necessarily a couple of SEs (foreground/background), and multiple spectral bands, we define extended SEs with:

- *shape* (*sh*) : the specific shape of the SE defined on the spatial domain,
- *band* (*b*) : the spectral band of the image assigned to the SE,
- *threshold* (*th*) : a threshold defined in \mathbb{N} (which may be avoided by using functional SEs),
- *type* (*ty*) : the type of the SE, either a greater bound (*G*) to be processed with a dilation or a lower bound (*L*) to be processed with an erosion.

The proposed model of extended SE includes spatial information like classical SE but also spectral information. SEs are defined to match a specific pattern, so parameters can be set from user problem knowledge or automatically. SEs are then used as follows for extraction of predefined templates.

First, we perform a *fitting* which will determine for each pixel if it matches the pattern defined by the set of SEs. Then, we perform a *valuation* which will assign a resulting value to each fitted pixel.

Fitting consists in checking for each pixel and for each SE if all pixel neighbours match the SE (not only spatially). If SE type is *greater* (respectively *lower*), SE threshold must be greater (resp. lower) than all neighbouring pixel (the neighbourhood being defined by the SE shape) value for the considered spectral band, as illustrated in Fig. 1.

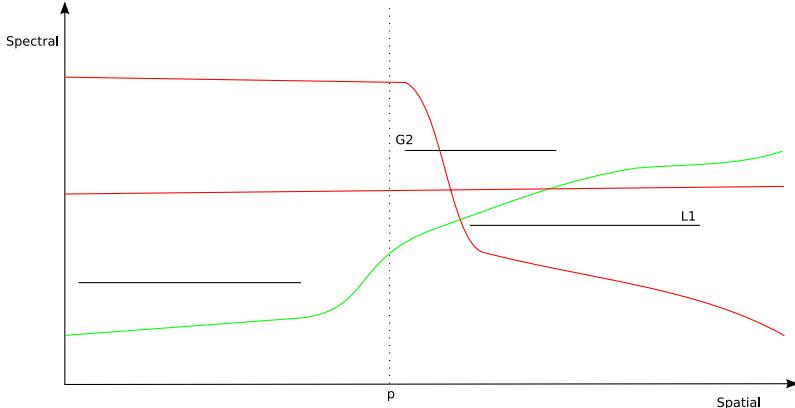


Fig. 1. Fitting example for a pixel p with a set of 3 linear SEs related to the same spectral band but with various shapes. We can observe that only the green curve is fitted and the red curves are not (G_1 and G_2 are greater SEs, L_1 is a lower SE).

The fitting is formally defined for a given SE s as:

$$\text{Fitting}_s(F)(p) = \begin{cases} \varepsilon_{s_{sh}}(F_{s_b})(p) \geq s_{th} & \text{if } s_{ty} = L \\ \delta_{s_{sh}}(F_{s_b})(p) \leq s_{th} & \text{if } s_{ty} = G \end{cases} \quad (5)$$

To be valued, a pixel must match all the SEs, i.e. global fitting is defined as follows:

$$\text{Fitting}_S(F)(p) = \bigcap_{s_i \in S} \text{Fitting}_{s_i}(F)(p) \quad (6)$$

where S is the set of cardinality $|S|$ of SEs defined by $S = \{s_i\}$.

Valuation consists in assigning a scalar value to all pixels. For non-fitted pixels, valuation process assigns the value 0 (similar to other HMTs, see section 2.2). For fitted pixels, the valuation is more complex and defined for each SE s by the following formula:

$$\text{Valuation}_s(F)(p) = \begin{cases} \frac{\varepsilon_{s_{sh}}(F_{s_b})(p) - s_{th}}{F_{s_b}^+ - s_{th}} & \text{if } s_{ty} = L \\ \frac{\delta_{s_{sh}}(F_{s_b})(p) - s_{th}}{F_{s_b}^- - s_{th}} & \text{if } s_{ty} = G \end{cases} \quad (7)$$

Global valuation is then defined as follows :

$$\text{Valuation}_S(F)(p) = \begin{cases} \frac{1}{|S|} \sum_{s_i \in S} \text{Valuation}_{s_i}(F)(p) & \text{if } \text{Fitting}_S(F)(p) \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

We can notice that MHMT results are normalized into $[0, 1]$ and do not depend anymore on the value range of the different spectral bands of the processed image. This specificity can be used or not, depending on the application (and its need for fuzzy or soft results).

Moreover, the proposed MHMT can also be applied on grey-level images. Indeed, a grey-level image is a multivalued image with a number of bands equal to one. The previous definitions can then be used relying on a single spectral band s_b .

3.2 Comparison with Related Work

MHMT offers several advantages compared to other HMT definitions. First, whereas Ronse's and Soille's HMTs use only spatial information through the shape of SEs (and their intensity profile with Ronse's HMT), MHMT also relies on spectral information with bands and thresholds. The combined use of spatial and spectral features allows extraction of well-defined templates in terms of spectral signature within a spatial configuration. Second, it can use more than two SEs which allows complex shapes to be matched (moreover SEs do not need to be disjoint anymore). Third, MHMT results are normalized in $[0, 1]$ which have two main interests: in one hand it can lead to fuzzy or soft interpretation of the results, in the other hand it allows comparison between a set of single MHMT results even if the spectral bands do not have the same value range. Finally, contrary to the HMT from Ronse or Soille which requires the two SEs to be fully applied on the image, our definition of MHMT allows a fast implementation. Indeed, the MHMT fitting needs each SE to be fitted: so, as soon as one SE is not fitted, processing the other SEs is useless. Furthermore, MHMT fitting compares erosion and dilation

results to a threshold, pixel values could directly be compared to threshold because of the same principle: as soon as one pixel value is lower than the threshold while processing an erosion with a lower SE, this SE is not fitted (making the processing of the rest of the SE useless). By simply extending Ronse's and Soille's definitions to multivalued space (cf. 2.3) using multivariate morphology [3], these interesting properties cannot be reached.

We will now give a practical illustration of the proposed MHMT for template matching in multivalued images, and focus on coastline extraction in multispectral images.

4 Application to Template Matching

We do not provide here a generic method for template matching by using MHMT, but we give some clues for elaborating methods adapted to the problem to solve and illustrate it by an application to coastline extraction.

4.1 MHMT-Based Template Matching

MHMT is a powerful tool for template matching, it can be used to build feature extraction methods. Due to its specificities, the set-up of MHMT-based template matching methods is quite simple and consists mainly in two steps (cf. figure 2).

The first step for the elaboration of feature extraction method is to formalize the problem and to wonder how the feature can be defined from both spatial and spectral points of view. Spatial part consists in definition of the shape and the position of each SE (related to its center), while spectral part is the introduction of expert knowledge about spectral response of feature to be extracted.

The second step concerns specificities of the problem to be solved, introduction of multiple SE's orientations if the sought template can have different orientations, or processing of MHMT at various scales if the template can have different sizes. These specification examples are obviously not exhaustive.

4.2 Example: Coastline Extraction

Coastline is typically defined as the border between sea and land. Its extraction is a well-known problem in remote sensing imagery [4]. Even if there exist several methods which give accurate results on low or medium resolution, few are relevant on very high spatial resolution (VHR) satellite imagery where a pixel represents an area lower than 5x5 meters square. Here, we define a method for coastline extraction using the MHMT.

The proposed MHMT has been used to build a coastline detection method for VHR images. Spatial definitions of SEs needed for coastline extraction are deduced from coastline definition. Coastline is defined as the border between sea and land, we defined two linear SEs which are both aligned and separated by only one pixel (S^1 and S^2 in figure 3). Using only these two SEs leads to extract all borders between water and land (e.g. lake border, river bank, etc...),

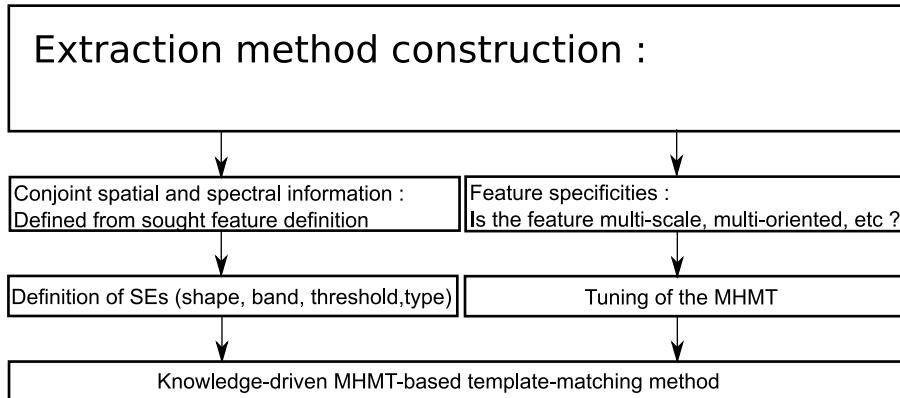


Fig. 2. Flowchart of MHMT-based method elaboration

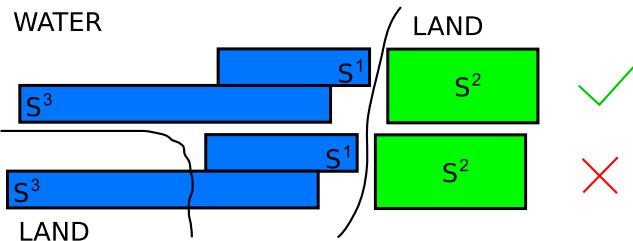


Fig. 3. Spatial definition of SEs used for coastline extraction with matching and unmatching conditions

so we have to add one more SE representing water further from the coastline (\$S^3\$ in figure 3). As coastline can have any orientation, MHMT is then applied with various directions. Now, we have defined the spatial part of our coastline extraction method.

The spectral parameters have been defined by a geographical expert, based on knowledge about spectral response of land (i.e. sand) or sea (i.e. water) in the area observed with a given satellite sensor. They are defined as follow :

$$\Omega = \begin{cases} s_{sh}^1 = 180 + 0, & s_b^1 = \text{NDVI}, s_{th}^1 = 0.5, & s_{ty}^1 = G \\ s_{sh}^2 = -180 + 0, & s_b^2 = \text{NDVI}, s_{th}^2 = 0.5, & s_{ty}^2 = L \\ s_{sh}^3 = 360 + 12, & s_b^3 = R, & s_{th}^3 = 0.05, s_{ty}^3 = L \end{cases} \quad (9)$$

where \$s_{sh} = i + j\$ means a line of length \$i\$ meters shifted of \$j\$ meters from the origin (minus representing opposite direction). R and NDVI denote normalized images respectively built from the red band and the normalized difference vegetation index.

Using only MHMT with parameters defined in (9) could lead to extract uncontinuous coastline due to satellite picture defaults, noise, etc. To improve results



Fig. 4. Coastline extraction on Normandy coast QuickBird image at resolution of 2.4m per pixel ((c)Digitalglobe)

on this specific application, we added a binary post-processing step to obtain a connected coastline. It consists in enlarging the initial thresholded MHMT result (all pixels with non null values are set to 1) to get a connected region and then compute and filter the morphological skeleton of this region to generate the final coastline result. The input of the skeleton algorithm, i.e. the connected region or mask build from the initial MHMT result, is obtained through a double threshold operator [5]. More precisely, another MHMT is applied with some more tolerant spectral parameters (i.e. thresholds) and is thresholded (with non null values set to 1) and filtered (with a morphological closing) to remove holes. Then we perform a geodesic reconstruction using the first MHMT as a marker and the second MHMT as a mask. On this reconstructed image is then extracted the skeleton considered as the final coastline result.

Fig. 4 illustrates the results obtained on a QuickBird image (2.4m/pixel spatial resolution) and shows the relevance of MHMT for the extraction of coastline in VHR imagery. Compared to existing approaches, the MHMT gives particular promising results for VHR imagery [6], due to the simultaneous usage of both spatial and spectral knowledge. In particular, MHMT-based method gives less false-positive pixels and is more accurate than existing methods [289]. Moreover, such an extraction could not been achieved with classical HMTs (such as those from Ronse or Soille) due to their inability to consider more than two shapes and to deal with spectral properties. Similarly, spectral approaches could not extract coastline as well. Indeed, spectral approaches can only determine borders between water and land regions, so they will also extract lake border and riverbank as coastline.

5 Conclusion

In this paper, a Hit-or-Miss Transform dedicated to multivalued images has been proposed. Since an image with a single band is a particular case of multivalued

images, this method can also be used on grey-level images. The proposed operator takes into account spectral properties without requiring a vectorial ordering. It relies on both spatial and spectral knowledge to define the set (of variable cardinality) of extended structuring elements (involving spectral features) and thus to extract specific templates. Its relevance for multivalued template matching has been underlined by solving the problem of coastline extraction in VHR remote sensing.

Future works will consist in using the MHMT in other application fields. To ensure a higher robustness to noise, a soft or fuzzy version of the MHMT operator will also be elaborated by replacing the global fitting and valuation steps by some more advanced merging techniques.

Acknowledgements

The authors thank Anne Puissant (LIV, CNRS / ULP) for the definition of the SE parameters used in the coastline extraction experiment and the ANR-JC ECOSGIL¹ project for the remote sensing data and financial support.

References

1. Serra, J.: *Image Analysis and Mathematical Morphology*. Academic Press, London (1982)
2. Naegel, B., Passat, N., Ronse, C.: Grey-level hit-or-miss transforms - part 1: Unified theory. *Pattern Recognition* 40(2), 635–647 (2007)
3. Aptoula., E., Lefèvre, S.: A comparative study on multivariate mathematical morphology. *Pattern Recognition* 40(11), 2914–2929 (2007)
4. Boak, E., Turner, I.: Shoreline definition and detection: a review. *Journal of Coastal Research* 21(4), 688–703 (2005)
5. Soille, P.: *Morphological Image Analysis: Principles and Applications*. Springer, Berlin (2003)
6. Puissant, A., Lefèvre, S., Weber, J.: Coastline extraction in vhr imagery using mathematical morphology with spatial and spectral knowledge. In: XXIth ISPRS Congress - Benjing (2008)
7. Heene, G., Gautama, S.: Optimisation of a coastline extraction algorithm for object-oriented matching of multisensor satellite imagery. In: IEEE International Geoscience and Remote Sensing Symposium, vol. 6, pp. 2632–2634 (2000)
8. Jishuang, Q., Chao, W.: A multi-threshold based morphological approach for extracting coastal line feature in remote sensing. In: ISPRS Commission I Symposium on Integrated Remote Sensing at the Global, Regional and Local Scale (2002)
9. Bagli, S., Soille, P.: Morphological automatic extraction of pan-european coastline from landsat etm+ images. In: Fifth International Symposium on GIS and Computer Cartography for Coastal Zone Management (COASTGIS) (2003)

¹ <http://ecosgil.u-strasbg.fr>

Blur Insensitive Texture Classification Using Local Phase Quantization

Ville Ojansivu and Janne Heikkilä

Machine Vision Group, Department of Electrical and Information Engineering,
University of Oulu, PO Box 4500, 90014, Finland
`{vpo,jth}@ee.oulu.fi`

Abstract. In this paper, we propose a new descriptor for texture classification that is robust to image blurring. The descriptor utilizes phase information computed locally in a window for every image position. The phases of the four low-frequency coefficients are decorrelated and uniformly quantized in an eight-dimensional space. A histogram of the resulting code words is created and used as a feature in texture classification. Ideally, the low-frequency phase components are shown to be invariant to centrally symmetric blur. Although this ideal invariance is not completely achieved due to the finite window size, the method is still highly insensitive to blur. Because only phase information is used, the method is also invariant to uniform illumination changes. According to our experiments, the classification accuracy of blurred texture images is much higher with the new method than with the well-known LBP or Gabor filter bank methods. Interestingly, it is also slightly better for textures that are not blurred.

1 Introduction

Natural surfaces usually exhibit some repetitive intensity variations or patterns that are generally referred to as texture. Analysis of texture information is important in machine vision, and it has numerous applications including surface inspection, medical image analysis, and remote sensing [1]. In some applications, image degradations may limit the applicability of the texture information. One class of degradation is blur due to motion, out of focus, or atmospheric turbulence. Because image deblurring is very difficult and introduces new artifacts, it is desirable to be able to analyze texture in a way that is insensitive to blur.

The focus of this paper is on blur insensitive texture classification. There are not many texture analysis methods that are considered to be insensitive to blurring. A blur robust descriptor based on color constancy was proposed in [2]. Also, blur invariant moments [3] or the modified Fourier phase [4] could be used in principle, but they are mainly intended for global object recognition, not local texture analysis.

In this paper, we propose a new blur insensitive texture classification method, which is based on quantized phase of the discrete Fourier transform (DFT) computed in local image windows, and it is called local phase quantization (LPQ).

The codes produced by the LPQ operator are insensitive to centrally symmetric blur, which includes motion, out of focus, and atmospheric turbulence blur [5]. The LPQ operator is applied to texture identification by computing it locally at every pixel location and presenting the resulting codes as a histogram. Generation of the codes and their histograms is similar to the LBP method [6]. Local frequency analysis, often referred to as signal processing methods, has also been used for texture analysis previously. For a review, see [7]. One of the best known methods uses a bank of Gabor filters and is based on magnitude information [8]. Phase information has been used in [9] and histograms have been used in conjunction with spectral information in [10]. Nevertheless, blur sensitivity has not been considered as a criterion when designing these operators.

First we introduce the conditions under which the DFT phase is invariant to blur in Sect. 2. Then, Sect. 3 proposes the LPQ operator. Section 4 contains experimental results and Sect. 5 presents conclusions.

2 Blur Invariance Using Fourier Transform Phase

In digital image processing, the discrete model for spatially invariant blurring of an original image $f(\mathbf{x})$ resulting in an observed image $g(\mathbf{x})$ can be expressed by a convolution [5], given by

$$g(\mathbf{x}) = (f * h)(\mathbf{x}) , \quad (1)$$

where $h(\mathbf{x})$ is the point spread function (PSF) of the blur, $*$ denotes 2-D convolution and \mathbf{x} is a vector of coordinates $[x, y]^T$. In the Fourier domain, this corresponds to

$$G(\mathbf{u}) = F(\mathbf{u}) \cdot H(\mathbf{u}) , \quad (2)$$

where $G(\mathbf{u})$, $F(\mathbf{u})$ and $H(\mathbf{u})$ are the discrete Fourier transforms (DFT) of the blurred image $g(\mathbf{x})$, the original image $f(\mathbf{x})$, and the PSF $h(\mathbf{x})$, respectively, and \mathbf{u} is a vector of coordinates $[u, v]^T$. We may separate the magnitude and phase parts of (2), resulting in

$$\begin{aligned} |G(\mathbf{u})| &= |F(\mathbf{u})| \cdot |H(\mathbf{u})| \quad \text{and} \\ \angle G(\mathbf{u}) &= \angle F(\mathbf{u}) + \angle H(\mathbf{u}) . \end{aligned} \quad (3)$$

If we assume that the blur PSF $h(\mathbf{x})$ is centrally symmetric, namely $h(\mathbf{x}) = h(-\mathbf{x})$, its Fourier transform is always real-valued, and as a consequence its phase is only a two-valued function, given by

$$\angle H(\mathbf{u}) = \begin{cases} 0 & \text{if } H(\mathbf{u}) \geq 0 \\ \pi & \text{if } H(\mathbf{u}) < 0 \end{cases} . \quad (4)$$

This means that

$$\angle G(\mathbf{u}) = \angle F(\mathbf{u}) \quad \text{for all } H(\mathbf{u}) \geq 0 . \quad (5)$$

In other words, the phase of the observed image $\angle G(\mathbf{u})$ at the frequencies, where $H(\mathbf{u})$ is positive, is invariant to centrally symmetric blur. In the case of ideal motion and out of focus blur, the cross-section of $h(\mathbf{x})$ is rectangular [5]. This results in a spectrum $H(\mathbf{u})$ of which cross-section is a sinc function containing also negative values. The values of $H(\mathbf{u})$ are always positive before the first zero crossing at frequency $\approx (\text{blur length})/(\text{sampling frequency})$ that satisfies (5). In the case of Gaussian PSF, which models atmospheric turbulence blur [6], $H(\mathbf{u})$ is also Gaussian with only positive values that always satisfy the condition (5).

In practice, blur invariance cannot be completely achieved because of the finite size of the observed images. The convolution of the ideal image with the blur PSF extends beyond the borders of the observed image so that part of the information is lost. When the extent of blur is large enough compared with the image size, this border effect becomes noticeable.

3 Local Phase Quantization for Texture Classification

3.1 Short-Term Fourier Transform

The local phase quantization (LPQ) method is based on the blur invariance property of the Fourier phase spectrum described in Sect. 2. It uses the local phase information extracted using the 2-D DFT or, more precisely, a short-term Fourier transform (STFT) computed over a rectangular M -by- M neighborhood $\mathcal{N}_{\mathbf{x}}$ at each pixel position \mathbf{x} of the image $f(\mathbf{x})$ defined by

$$F(\mathbf{u}, \mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{N}_{\mathbf{x}}} f(\mathbf{x} - \mathbf{y}) e^{-j2\pi \mathbf{u}^T \mathbf{y}} = \mathbf{w}_{\mathbf{u}}^T \mathbf{f}_{\mathbf{x}} , \quad (6)$$

where $\mathbf{w}_{\mathbf{u}}$ is the basis vector of the 2-D DFT at frequency \mathbf{u} , and $\mathbf{f}_{\mathbf{x}}$ is another vector containing all M^2 image samples from $\mathcal{N}_{\mathbf{x}}$.

As it can be noticed from (6), an efficient way of implementing the STFT is to use 2-D convolutions $f(\mathbf{x}) * e^{-2\pi j \mathbf{u}^T \mathbf{x}}$ for all \mathbf{u} . Since the basis functions are separable, computation can be performed using 1-D convolutions for the rows and columns successively.

In LPQ only four complex coefficients are considered, corresponding to 2-D frequencies $\mathbf{u}_1 = [a, 0]^T$, $\mathbf{u}_2 = [0, a]^T$, $\mathbf{u}_3 = [a, a]^T$, and $\mathbf{u}_4 = [a, -a]^T$, where a is a scalar frequency below the first zero crossing of $H(\mathbf{u})$ that satisfies the condition (5). Let

$$\mathbf{F}_{\mathbf{x}}^c = [F(\mathbf{u}_1, \mathbf{x}), F(\mathbf{u}_2, \mathbf{x}), F(\mathbf{u}_3, \mathbf{x}), F(\mathbf{u}_4, \mathbf{x})] , \quad \text{and} \quad (7)$$

$$\mathbf{F}_{\mathbf{x}} = [\text{Re}\{\mathbf{F}_{\mathbf{x}}^c\}, \text{Im}\{\mathbf{F}_{\mathbf{x}}^c\}]^T , \quad (8)$$

where $\text{Re}\{\cdot\}$ and $\text{Im}\{\cdot\}$ return real and imaginary parts of a complex number, respectively. The corresponding 8-by- M^2 transformation matrix is

$$\mathbf{W} = [\text{Re}\{\mathbf{w}_{\mathbf{u}_1}, \mathbf{w}_{\mathbf{u}_2}, \mathbf{w}_{\mathbf{u}_3}, \mathbf{w}_{\mathbf{u}_4}\}, \text{Im}\{\mathbf{w}_{\mathbf{u}_1}, \mathbf{w}_{\mathbf{u}_2}, \mathbf{w}_{\mathbf{u}_3}, \mathbf{w}_{\mathbf{u}_4}\}]^T , \quad (9)$$

so that

$$\mathbf{F}_{\mathbf{x}} = \mathbf{W} \mathbf{f}_{\mathbf{x}} . \quad (10)$$

3.2 Statistical Analysis of the Coefficients

Let us assume that the image function $f(\mathbf{x})$ is a result of a first-order Markov process, where the correlation coefficient between adjacent pixel values is ρ , and the variance of each sample is σ^2 . Without a loss of generality we can assume that $\sigma^2 = 1$. As a result, the covariance between positions \mathbf{x}_i and \mathbf{x}_j becomes

$$\sigma_{ij} = \rho^{\|\mathbf{x}_i - \mathbf{x}_j\|}, \quad (11)$$

where $\|\cdot\|$ denotes L_2 norm, and the covariance matrix of all M samples in $\mathcal{N}_{\mathbf{x}}$ can be expressed by

$$\mathbf{C} = \begin{bmatrix} 1 & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{21} & 1 & \cdots & \sigma_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{M1} & \sigma_{M2} & \cdots & 1 \end{bmatrix}. \quad (12)$$

Hence, the covariance matrix of the transform coefficient vector $\mathbf{F}_{\mathbf{x}}$ can be obtained from

$$\mathbf{D} = \mathbf{W} \mathbf{C} \mathbf{W}^T. \quad (13)$$

One can easily notice that \mathbf{D} is not a diagonal matrix for $\rho > 0$, meaning that the coefficients are correlating.

3.3 Decorrelation and Quantization

Before quantization the coefficients are decorrelated, because it can be shown that the information is maximally preserved in scalar quantization if the samples to be quantized are statistically independent. Assuming Gaussian distribution, independence can be achieved using a whitening transform

$$\mathbf{G}_{\mathbf{x}} = \mathbf{V}^T \mathbf{F}_{\mathbf{x}}, \quad (14)$$

where \mathbf{V} is an orthonormal matrix derived from the singular value decomposition (SVD) of the matrix \mathbf{D} that is

$$\mathbf{D} = \mathbf{U} \Sigma \mathbf{V}^T. \quad (15)$$

Notice, that \mathbf{V} can be solved in advance for a fixed value of ρ .

Next, $\mathbf{G}_{\mathbf{x}}$ is computed for all image positions, i.e., $\mathbf{x} \in \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$, and the resulting vectors are quantized using a simple scalar quantizer

$$q_j = \begin{cases} 1, & \text{if } g_j \geq 0 \\ 0, & \text{otherwise} \end{cases}, \quad (16)$$

where g_j is the j th component of $\mathbf{G}_{\mathbf{x}}$. The quantized coefficients are represented as integer values between 0-255 using binary coding

$$b = \sum_{j=1}^8 q_j 2^{j-1}. \quad (17)$$

Finally, a histogram of these integer values from all image positions is composed and used as a 256-dimensional feature vector in classification.

The resulting integers b are invariant to centrally symmetric blur provided that the window \mathcal{N}_x is infinitely large and the frequency spectrum of the blur PSF is positive at the sample locations $\mathbf{u}_1 - \mathbf{u}_4$. The second condition is easily met if a is sufficiently small. However, the first condition cannot be fulfilled in practice, and therefore, complete invariance is not achieved, but as shown in the experiments, even a relatively small neighborhood is enough for robustness to reasonable extents of blur.

Decorrelation and quantization do not have any effect on the blur invariance property. In the whitening transform the coefficient vectors are subject to an eight-dimensional rotation that only causes a uniform phase shift to all vectors. In quantization the eight-dimensional space is divided into 256 hypercubes, and the assignment of a vector to one of these hypercubes depends only on the phase information.

4 Experiments

In the experiments, we measured the performance of our LPQ method in the classification of sharp as well as blurred textures. The correlation coefficient was selected to be $\rho = 0.9$ in all the experiments. As test material we used the applicable test suites of the Outex texture image database¹ [1]. For comparison, we also did the same experiments with two other widely known texture classification methods: local binary pattern (LBP) method² [2][6] and a method based on Gabor filter banks³ [8]. We used the Matlab implementations of these reference methods, which can be found on the Internet. Both methods have also been used previously in conjunction with the Outex texture database [1].

All three test suites of the Outex texture database used in our experiments, *Outex_TC_00000-00002*, contained images from 24 texture classes and had 100 different test cases that divided the images into training and test sets differently. Test suites with a larger number are more challenging, as they contain more and smaller images. In the experiments, we used a k-nearest neighbor (k-NN) classifier, which was trained and tested using the appropriate sets of the images. The value of k was 1, 3 or 15 for test suites *Outex_TC_00000-00002*, respectively. We used the Chi square distance measure for the LPQ and LBP histograms. For the Gabor features we used the distance measure proposed in [8]. Notation $LBP_{P,R}$ means LBP with P samples at radius R . In classical LBP $P = 8$, which results in a code with values in the range $\{0, \dots, 255\}$, similar to LPQ. Notation LPQ_R means LPQ of a spatial window with dimensions $M = 2R + 1$. A larger radius R for LPQ and LBP, which provides the comparable spatial extent of the operators, gives better results for blurred textures, but too large radius deteriorates the classification results for sharp textures. The frequency parameter used for LPQ was $a = 1/M$, which is the lowest non-zero frequency.

¹ <http://www.outex.oulu.fi/>

² http://www.ee.oulu.fi/mvg/page/lbp_matlab/

³ <http://vision.ece.ucsb.edu/texture/software/>

Table 1. Texture classification accuracies of the non-whitened LPQ, LPQ, LBP and Gabor methods in the first experiment for *Outex_TC_00002* test suite

	LPQ ₁ nw	LPQ ₁	LBP _{8,1}	Gabor
Accuracy	88.0 %	93.6 %	90.2 %	90.2 %

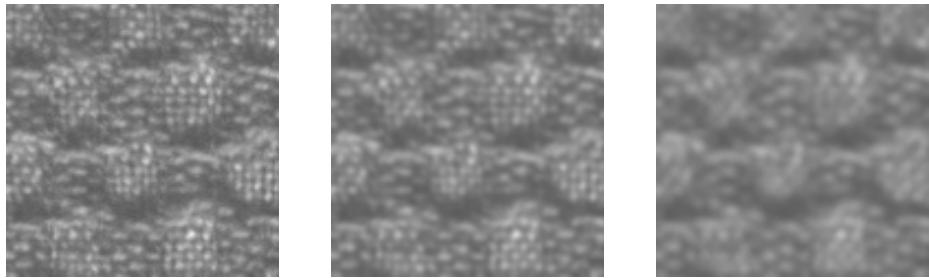


Fig. 1. An example of the texture images used in the second experiment (left). Circularly blurred versions of the same image with blur radii one (middle) and two (right).

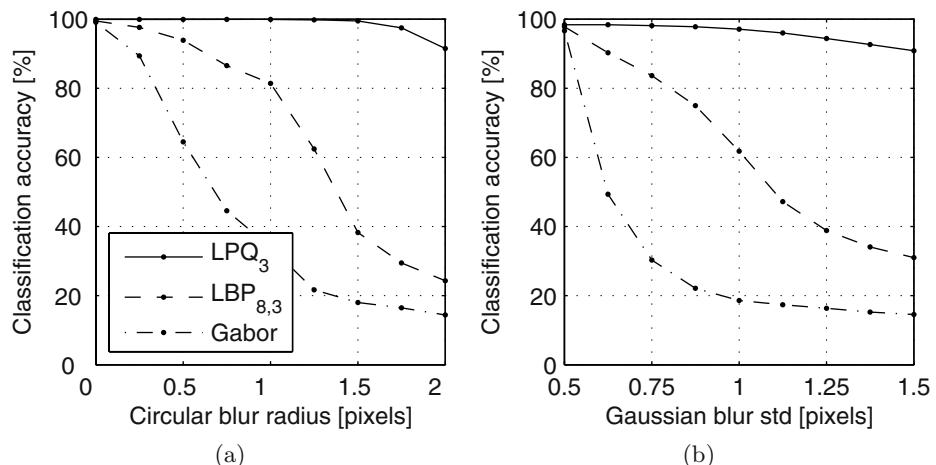


Fig. 2. Classification accuracies of the LPQ, LBP and Gabor methods for (a) circularly blurred textures of test suite *Outex_TC_00000* and for (b) Gaussian blurred textures of test suite *Outex_TC_00001*

In the first experiment, we tested the classification performance of the methods for the sharp texture images of the challenging *Outex_TC_00002* test suite. The test suite includes 8832 images of size 32×32 ; hence, 368 images per class. We also included the result for non-whitened (nw) LPQ to demonstrate the effect of the whitening transform. We used LBP_{8,1} and LPQ₁, which results in the basic

forms of these operators, and the Gabor method. The classification accuracy as percentages is shown in Table II.

As can be seen, the whitening improves the performance of the LPQ method significantly and the whitened LPQ gives the best score. The success of LPQ also for sharp images was a little surprising, but it was verified also in the other experiments. For some reason the result of the Gabor method for the same test suite was two percentages better in the experiments in III, where the Gabor method gave the best result. Nevertheless, the result of LPQ is still the best.

In the second and third experiment, we tested the three texture classification methods in the case of blurred textures, which is the main theme of this paper. In the second experiment we used the test suite *Outex_TC_00000*, which includes 480 images of size 128×128 , 20 images per class. The classifier was trained using the sharp images, but the test images were artificially blurred using circular and flat PSF, which mimics the out of focus blur [5]. The blur radius was $\{0, 0.25, \dots, 2\}$. Figure 1 shows three examples of one texture image with blur radii 0, 1, and 2, respectively. We used LPQ and LBP operators with various values of R , but value $R = 3$ seemed to give the best trade-off for different extents of blur. The results for these LPQ_3 , $\text{LBP}_{8,3}$, and Gabor methods are shown in Fig. 2(a). As can be seen from the diagram, the LPQ method is very tolerant of blur, while the Gabor method performs worst. Even very small blur deteriorates the results of all but the LPQ methods. We also tried a modification of $\text{LBP}_{8,3}$, namely $\text{LBP}_{16,3}^{u2}$ [6], which uses 16 samples, but the result was only 0.2 percent better for sharp images and worse for blurred images. It is remarkable that the result for LPQ also for sharp images (Blur radius = 0) was the best for any values of R .

In the third experiment, we used the *Outex_TC_00001* test suite, which includes 2112 images of size 64×64 and thus 88 images per class. Now, the artificial blur had Gaussian PSF with standard deviation in the range $\{0.5, 0.75, \dots, 1.5\}$, which mimics the blur caused, for example, by atmospheric turbulence [5]. Otherwise, the experiment was similar to the second experiment. Again, LPQ_3 and $\text{LBP}_{8,3}$ offered the best trade-off for different extents of blur; therefore, the results of these operators with the Gabor method are shown in Fig. 2(b). The alternative $\text{LBP}_{16,3}^{u2}$ operator did not improve the results. As can be seen, the results of Fig. 2(a) and Fig. 2(b) are quite similar, except that the test suite is a bit more challenging in the latter. The LPQ_3 is again the best option at any blur level, while the Gabor method is the worst. For all values of R the result of LPQ was also the best for the smallest blur (Blur std = 0.5).

5 Conclusions

In this paper, we proposed a new LPQ texture analysis method that operates on the Fourier phase computed locally for a window in every image position. The phases of the four low-frequency coefficients are uniformly quantized into one of 256 hypercubes in eight-dimensional space, which results in an 8-bit code. These LPQ codes for all image pixel neighborhoods are collected into a histogram,

which describes the texture and can be used for classification. The phases of the low-frequency components are shown to be ideally invariant to centrally symmetric blur. Although, the invariance is disturbed by the finite-sized image windows, the method is still very tolerant of blur. Because only phase information is used, the method is also invariant to uniform illumination changes.

The proposed method was compared with two well-known texture analysis operators, the LBP and a Gabor filter bank based method. The results of the texture classification experiments on the Outex texture database show that the LPQ method tolerates significantly more blurring than other methods. In addition to that, LPQ also gave slightly better results for sharp texture images.

References

1. Tuceryan, M., Jain, A.K.: Texture analysis. In: Chen, C.H., Pau, L.F., Wang, P.S.P. (eds.) *The Handbook of Pattern Recognition and Computer Vision*, pp. 207–248. World Scientific Publishing Co, Singapore (1998)
2. van de Weijer, J., Schmid, C.: Blur robust and color constant image description. In: Proc. IEEE International Conference on Image Processing (ICIP 2006), Atlanta, Georgia, October 2006, pp. 993–996 (2006)
3. Flusser, J., Suk, T.: Degraded image analysis: An invariant approach. *IEEE Trans. Pattern Anal. Machine Intell.* 20(6), 590–603 (1998)
4. Ojansivu, V., Heikkilä, J.: A method for blur and similarity transform invariant object recognition. In: Proc. International Conference on Image Analysis and Processing (ICIAP 2007), Modena, Italy, September 2007, pp. 583–588 (2007)
5. Banham, M.R., Katsaggelos, A.K.: Digital image restoration. *IEEE Signal Processing Mag.* 14(2), 24–41 (1997)
6. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Machine Intell.* 24(7), 971–987 (2002)
7. Randen, T., Husøy, J.H.: Filtering for texture classification: A comparative study. *IEEE Trans. Pattern Anal. Machine Intell.* 21(4), 291–310 (1999)
8. Manjunathi, B.S., Ma, W.Y.: Texture features for browsing and retrieval of image data. *IEEE Trans. Pattern Anal. Machine Intell.* 18(8), 837–842 (1996)
9. Vo, A.P., Oraintara, S., Nguyen, T.T.: Using phase and magnitude information of the complex directional filter bank for texture image retrieval. In: Proc. IEEE International Conference on Image Processing (ICIP 2007), San Antonio, Texas, September 2007, pp. 61–64 (2007)
10. Xiuwen, L., DeLiang, W.: Texture classification using spectral histograms. *IEEE Trans. Image Processing* 12(6), 661–670 (2003)
11. Ojala, T., Mäenpää, T., Pietikäinen, M., Viertola, J., Kyllönen, J., Huovinen, S.: Outex - new framework for empirical evaluation of texture analysis algorithms. In: Proc. 16th International Conference on Pattern Recognition (ICPR 2002), August 2002, pp. 701–706 (2002)
12. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distribution. *Pattern Recognition* 29(1), 51–59 (1996)

An Algorithm to Determine the Chromaticity Under Non-uniform Illuminant

Sivalogeswaran Ratnasingam and Steve Collins

Department of Engineering Science,
University of Oxford, OX1 3PJ, Oxford, United Kingdom
{siva, collins}@robots.ox.ac.uk

Abstract. Colour based object recognition is a difficult problem because of the effect of scene illuminant and geometry on the captured image. In this paper the ability of an algorithm proposed by Finlayson and Drew [1] to separate similar colours is assessed. A new variant of this algorithm is then proposed that results in a slight improvement in performance. A significant performance improvement is achieved by optimising the characteristics of the sensors that are used to acquire the data for this algorithm. This optimisation process results in several combinations of sensors and associated data projections that have a comparable performance when required to distinguish between similar colours. Since this performance is comparable to that of the human visual system it is suggested that with the correct sensors this algorithm is capable of obtaining useful chromaticity information under varying illumination conditions.

Keywords: color based recognition, chromaticity constancy, colour indexing.

1 Introduction

Digital cameras have been an undoubted success in the consumer market in the last few years. However, compared to human visual system these cameras have two limitations, a limited dynamic range and variability in the apparent colour of objects. The first of these limitations causes saturation when the dynamic range of the scene is larger than that of the camera. The second limitation is subtler and means that, unlike human vision, colour information can not always be reliably used in artificial object recognition systems.

An early approach for recognizing objects based on colour was developed by Swain and Ballard [2] called ‘colour indexing’. This work is a significant contribution to colour based object recognition, but this method does not address the issue of variation of scene geometry and illumination. To solve this problem Funt and Finlayson[3] created colour indices using the ratio of responses from neighbour pixels. An alternative approach proposed by Berwick and Lee [4] used a diagonal transform to model the sensor response and a translation to model the effect of variations in the illuminant spectrum. This general approach was followed by Finlayson and Drew, who proposed an algorithm using the responses from four sensors to first compensate for changes in scene geometry and illuminant intensity, followed by a projection into a space in which the effect of changing the illuminant spectrum was minimized.

In this paper, Finlayson and Drew's algorithm [1] and a flexible method for removing illuminant effect are optimised to achieve better chromaticity constancy. The optimisation is done using a steepest descend algorithm and the performance of variations of both algorithms (normalizing by one of the channel response and by geometric mean of all four responses) is also investigated. Both algorithms map different colours into a 2-dimensional space, invariant to scene geometry and illuminant. The 2 dimensional spaces formed by these algorithms show a smooth spatial variation of colours across the space. Initial results suggest that in this space, colours that are separated by between 6 units in the CIELab space could be identified uniquely.

2 Algorithm

Differences between the spectra of different illuminants can cause significant changes between the receptor responses to the same surface when illuminated by the different illuminants. The aim of the algorithms is to extract an invariant representation of each colour under different illuminants.

The illuminants data used to assess these algorithms were obtained from CIE standard daylights [5] and the surfaces that were used were from the Munsell data set [6]. For each of the 1269 surfaces in this data set the reflectance spectra, measured using a spectrometer at 1 nm intervals, and the different colour co-ordinates, when viewed under a D65 illuminant, were available. For this study the sensitivity of the sensors were modeled using a Lorentzian function characterized by its full width at half the maximum response.

Finlayson and Drew. [1] started with the logarithm of each of four sensor responses. Then either one of these logarithmic sensor responses or the geometric mean of all four was used as a normalizing channel to compensate for possible changes in geometry and the brightness of the illuminant [7] [8]. When a single logarithmic response was used as the normalizing channel this removes the dependency on illuminant intensity and scene geometry while reducing the dimension of the space of possible responses by one. Alternatively, when the geometric mean is used as the normalizing channel one of the normalized channels is discarded to achieve the same effect. In either case the result is a 3 dimensional space of possible normalized response ratios. Within this space each colour follows a trajectory as the spectrum of the illuminant varies. To remove this variation, the 3 dimensional space is projected into a 2 dimensional space in which the variations due to changing illuminant spectrum are minimized.

Initially the performance of the algorithm proposed by Finlayson and Drew was investigated when used with four sensors whose responses are characterized by the parameters, given in table 1, chosen to give an even spread of peak sensitivities in the wavelength range 400nm to 700nm. The data used to assess these algorithms was

Table 1. Initial sensor parameters

Sensor ID	1	2	3	4
Peak position (nm)	437.5	512.5	587.5	662.5
Width (nm)	60	60	60	60

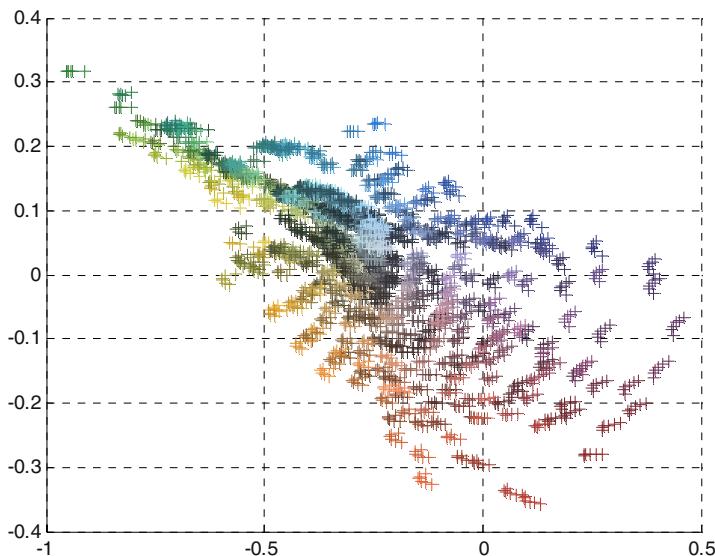


Fig. 1. 671 Munsell colours mapped on the two dimensional chromaticity space formed by the sensors given in table 1

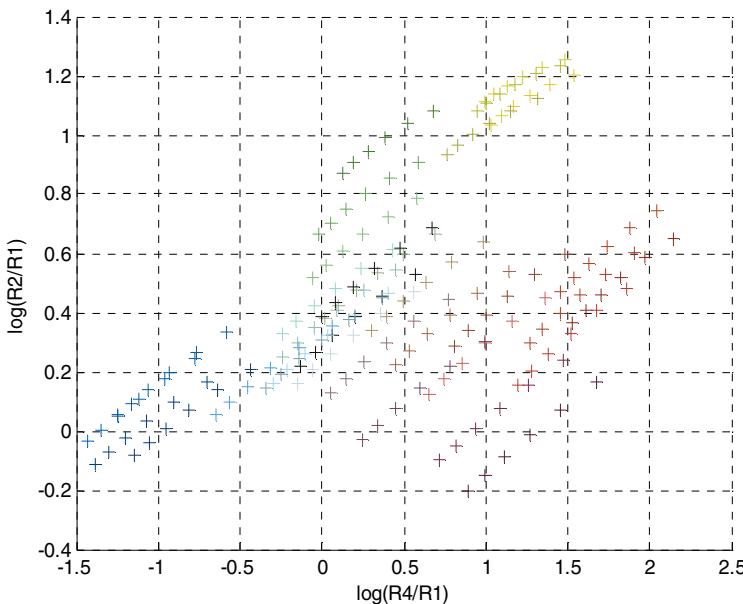
obtained by integrating the triple product of the reflectance spectra of each different colour, one of six CIE standard daylight illuminants with a colour temperature between 4000 K and 6500 K and a sensor response function over the wavelength range between 400nm and 700nm. The results in figure 1 show that the proposed algorithm creates a two dimensional space with similar colours projected into neighbouring areas. This suggests that noise in the data will not lead to a large error in the interpretation of the surface chromaticity.

This is a promising result. However, the most important assessment criterion for the algorithm is the ability to use its output to reliably distinguish colours. One colour space that has been defined to measure the perceptual differences between colours is the CIELab space. Ideally there is a just noticeable difference between two colours that are separated by a unit distance in CIELab space. However, there are various descriptions of the quality of colour matching associated with different distances in the space. For example matching colours between 3 and 6 units have been described as ‘good quality’ [9] or ‘acceptable’ [10] matching. To compare the performance of the algorithm to the human visual system, pairs of colours from the data set were identified that are separated by known distances in CIELab space.

As 6-CIELab unit is acceptable quality of colour reproduction, the algorithm was tested using 135 pairs of colours with the members of each pair chosen so that they are separated by 6-CIELab units. A close inspection of the final two dimensional space shows that each colour is projected into a small area within this space. A pair of colours was deemed to be distinguishable or separable from each other if the minimum distance between the two small areas was larger than the average of the largest dimensions of the two small areas. The results in Table 2 for the algorithm,

Table 2. Separability of colour pairs

Normalising sensor	Colour separability (%)
1	80
2	85
3	75
4	80
Geometric mean and discard 1	78
Geometric mean and discard 2	80
Geometric mean and discard 3	75
Geometric mean and discard 4	80

**Fig. 2.** Two dimensional Log channel ratio space shows the variation of different colours under six CIE standard day light illuminants

with the different options for normalization, suggest that the algorithm performs well in this task, however, there is some room for improvement.

One possible approach for improving the algorithm is to replace the direct projection from a three dimensional space after normalization to the two dimensional chromaticity space with a more flexible two step projection procedure. In this alternative approach, as before the first step is to use either one of the sensor responses or the geometric mean of all four responses to compensate for possible changes in geometry of the scene and the brightness of the scene illuminant. In the latter case one of the normalized responses is then again discarded. The resulting three ratios of responses are then used to create 2 dimensional spaces. A typical 2 dimensional space is shown

in figure 2. The variability in this 2 dimensional space is caused by the changes in the illuminant. Feature extraction based on a Karhunen-Loëve (KL) transform is therefore applied to each of these spaces to extract the feature that contains the minimum illuminant induced variability. Figure 3 shows the feature extracted from the data in figure 2 for a range of Munsell colours. These results show that this transformation has resulted in a feature that is almost independent of the illuminant. However, it is clear that some very different colors, such as samples 9, 11 and 15, in figure 3 have the same feature value and would therefore be confused.

This confusion is avoided by using the features obtained from an equivalent KL transformation applied to the two additional 2-dimensional spaces created from the other pairs of response ratios (as shown in figure 4). The three features obtained by this process can be used to create a 3 dimensional space that is largely independent of the illuminant and scene geometry. Any variations in this space are therefore dominated by differences between colours. In this space a feature extraction based on KL transform is therefore applied to extract the two features that contain the most variability. These two features are then used to form the chromaticity space.

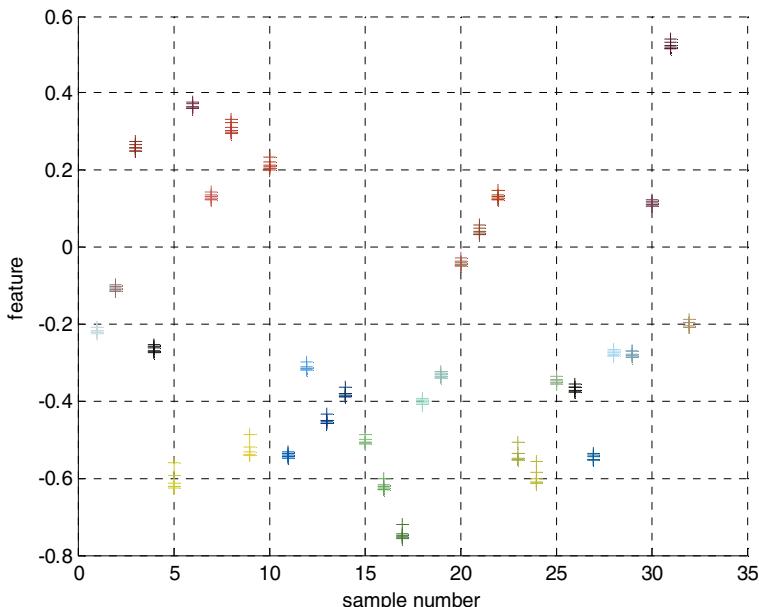


Fig. 3. Extracted chromaticity features from the two dimensional space shown in figure 2

The performance of this alternative procedure for obtaining a chromaticity space was assessed using the same sensor positions and test data as used to generate the results in table 2. Although there appears to be a slight improvement in performance this improvement is only a fraction of the possible improvement.

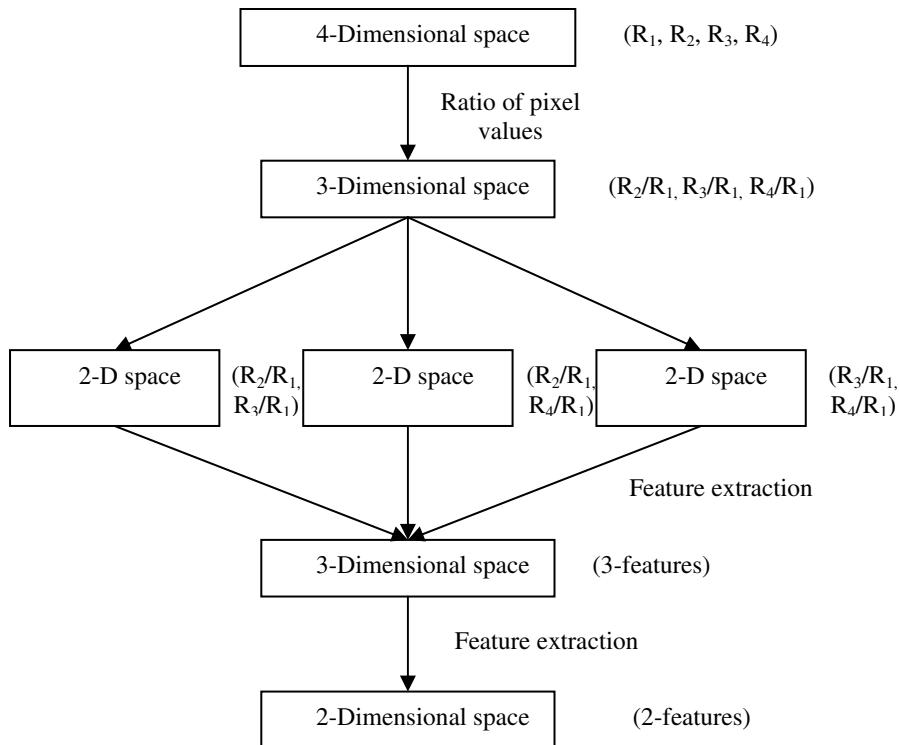


Fig. 4. The proposed method of obtaining a chromaticity space when one of the sensor responses is used to remove the effects of the lighting conditions

Table 3. Test results of the initial space using 6-CIELab test pairs

Normalising sensor	Colour separability (%)
1	85
2	80
3	80
4	80
Geometric mean and discard 1	80
Geometric mean and discard 2	80
Geometric mean and discard 3	77
Geometric mean and discard 4	73

3 Sensor Optimisation

The algorithms described in section 2 have several parameters, including the sensor characteristics and the projections into the 2 dimensional space. A closer inspection of

results such as those in figure 1 shows that although ideally each colour should be projected to a unique point in the space, in fact each colour is projected to a small area. This suggests that it might be possible to improve the separability of colours by optimising the sensor characteristics to minimise the projected area of each colour. To investigate the possibility of optimising the sensor characteristics, a steepest descent algorithm has been used to vary the position and width of each of the four sensors in a way that improves the separability of colours in the chromaticity space.

For this optimisation, 57 pairs of colours with the members of each pair chosen so that they were separated in CIELab space by between 2.95 and 3.05 units was used as the training data. There is a possibility that if the parameters are chosen to minimize the average size of the cluster of points from each colour it would merge the different colours. The error measure that was minimized by the steepest descent algorithm was therefore the average ratio of largest dimension of the small area representing each colour to the minimum distance between each pair of colours. With each of the four sensors characterized by a wavelength of peak sensitivity and a sensitivity width at half maximum, there are eight parameters that can be varied independently.

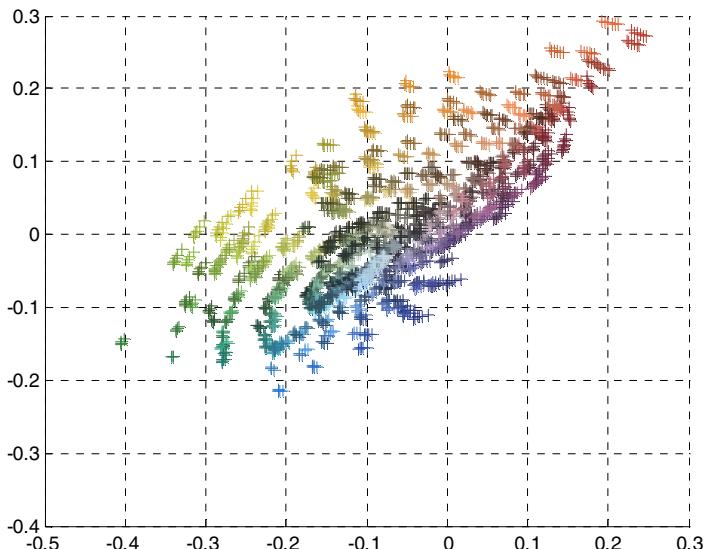


Fig. 5. The two dimensional space formed with a set of optimised sensors obtained when starting from the parameters given in table 1

The typical result in figure 5 is the optimised space corresponding to the initial space shown in figure 1. Again similar colours are neighbours in this space. However, a comparison of the two figures show that the result of this optimisation process is to shrink the area in the two dimensional space covered by the Munsell colours. The results in table 4 show that for all methods of normalisation and projection, when starting with the sensor parameters from table 1, the optimisation lead to a reduction in the area covered by the Munsell colours.

Table 4. Comparison of initial and final chromaticity spaces

Normalising channel	Area of the initial space	Area of the final space	Initial capacity	Final capacity
Results for direct projection				
1	1.1	0.40	0.97e4	3.0 e4
2	2.1	0.18	1.1 e4	1.6 e4
3	1.5	0.26	0.82e4	1.2 e4
4	1.2	0.37	0.93e4	1.1 e4
Geometric mean (discard 1)	0.6	0.14	0.86e4	3.3 e4
Geometric mean (discard 2)	0.4	0.15	0.86e4	0.86e4
Geometric mean (discard 3)	0.6	0.53	1.1 e4	1.1 e4
Geometric mean (discard 4)	0.7	0.34	1.0 e4	1.3 e4
Results for indirect projection				
1	1.7	0.36	1.0 e4	4.0 e4
2	3.2	0.85	1.1 e4	2.4 e4
3	2.1	0.31	0.78e4	0.88e4
4	1.6	0.09	0.87e4	3.9 e4
Geometric mean (discard 1)	1.0	0.47	0.94e4	1.4 e4
Geometric mean (discard 2)	0.6	0.33	0.83e4	3.2 e4
Geometric mean (discard 3)	0.8	0.35	1.1 e4	1.8 e4
Geometric mean (discard 4)	0.9	0.10	1.0 e4	0.80e4

Table 5. Comparison of colour separability of colours that differ by 6 CIELab units

Normalizing channel response	Colour separability of initial space (%)	Colour separability of optimised space (%)
Results for direct projection		
1	80	94
2	85	90
3	75	90
4	80	80
Geometric mean (discard 1)	78	90
Geometric mean (discard 2)	80	88
Geometric mean (discard 3)	75	81
Geometric mean (discard 4)	80	85
Results for indirect projection		
1	85	95
2	80	92
3	80	85
4	80	92
Geometric mean (discard 1)	80	93
Geometric mean (discard 2)	80	94
Geometric mean (discard 3)	77	90
Geometric mean (discard 4)	73	88

Although the reduction in the area of the projected space is unexpected the optimisation process was designed to maximise the distinguishability of colours. It is possible that this smaller space is better than the previous space if the projected area of each colour has reduced by a larger fraction than the area covered by all the colours. In table 4, the area of the chromaticity space has been estimated using the area of a smallest ellipse which encloses all the 1269 munsell colours. As the area covered by a typical individual colour is curvilinear in shape the area of each colour cluster is calculated by fitting smallest ellipse around each colour. The average area occupied by a colour is then obtained by averaging the areas of all the 1269 colours. The number of colours that can be accommodated within the project space was then estimated by finding the ratio between the area of chromaticity space to this average area. Table 4 shows that although the areas of the two dimensional spaces have been decreased by optimising the sensor characteristics, in most cases there are potentially more distinguishable colours in the chromaticity spaces after optimisation.

The results in table 5 show that as required the optimisation process has increased the percentage of colour pairs that can be distinguished after projection onto the chromaticity space. Furthermore, the performance of the better combinations of sensors and projection methods are so good and so similar that it is impossible to determine which of them is the best. However, from these results it is clear that the performance of these algorithms has been significantly improved by optimisation of the spectral sensitivities of the sensors.

4 Conclusions

An algorithm that extracts a two dimensional representation of the chromaticity of different colours has been investigated. The algorithm uses the outputs from four sensors to discount the effect of scene illuminant and geometry and since it processes information at the pixel level it is expected to work well in non-uniformly illuminated scenes.

The ultimate performance of the algorithm will be application dependant. However, it has been suggested that its general ability to distinguish colour chromaticity under varying illuminants can be assessed using Munsell colours with known separations in CIELab space. With uniformly spaced sensors the performance of the algorithm as proposed by Finlayson and Drew was encouraging but not perfect. A variant of this algorithm with more parameters was therefore investigated with a marginal improvement in performance. A method to change the sensor characteristics to optimise the performance of the algorithm when required to distinguish similar colours was therefore proposed. Several variants of the algorithm could then distinguish more than 90% of the pairs of perceptually ‘well-matched’ colours. Since the performance of several variants of the algorithms were indistinguishable other factors, such as the availability of sensors with the appropriate characteristics will determine the final choice between the different variants of the algorithm.

References

1. Finlayson, G.D., Drew, M.S.: 4-sensor camera calibration for image representation invariant to shading, shadows, lighting, and specularities. In: ICCV 2001: IEEE International Conference on Computer Vision, pp. 473–480 (2001)
2. Swain, M.J., Ballard, D.H.: Color Indexing. International Journal of Computer Vision 7, 11–32 (1991)
3. Funt, B.V., Finlayson, G.D.: Color constant color indexing. IEEE transactions on Pattern analysis and Machine Intelligence, 522–529 (1995)
4. Berwick, D., Lee, S.W.: A Chromaticity Space for Specularity, Illumination color- and Illumination Pose-Invariant 3-D Object Recognition. In: Proc. Int'l Conf. Computer Vision, pp. 165–170 (1998)
5. Colorpro Communications, <http://www.colorpro.com/info/data/daylites.html>
6. Database – Munsell Colours Matt, <ftp://ftp.cs.joensuu.fi/pub/color/spectra/mspec/>
7. Finlayson, G.D., Hordley, S.D.: Color constancy at a pixel. J. Opt. Soc. Am. A 18, 253–264 (2001)
8. Finlayson, G.D., Schiele, B., Crowley, J.L.: Comprehensive Colour Image Normalization. In: Fifth European Conference on Computer Vision, Freiburg, Germany, June 1998, pp. 475–490 (1998)
9. Abrardo, A., Cappellini, V., Cappellini, M., Mecocci, A.: Art-works colour calibration using the.VASARI scanner. In: Proceedings of IS&T and SID's 4th Color Imaging Conference: Color Science, Systems and Applications, pp. 94–97, Scottsdale, Arizona (1996)
10. Hardeberg, J.Y.: Acquisition and reproduction of color images: colorimetric and multispectral approaches. Ph.D. dissertation (Ecole Nationale Supérieure des Télécommunications, Paris (1999))

Transients Detection in the Time-Scale Domain

V. Bruni and D. Vitulano*

Istituto per le Applicazioni del Calcolo - C.N.R., Viale del Policlinico 137,
00161 Rome Italy
`{bruni,vitulano}@iac.rm.cnr.it`

Abstract. In this paper a novel model for transients detection in piecewise stationary signals is presented. A hybrid representation is assumed for the signal and the different behavior of each component (stationary, transient and stochastic) in the time-scale plane is exploited. Experimental results on both shape contours, described by a differential chain code, and audio signals show the generality of the proposed model.

Keywords: Transients, chain code, shape analysis, audio signal, wavelets.

1 Introduction

A classical problem of signal and image processing is the detection of signal significant components, such as edges and textures for images or transients and harmonic parts for audio signals. A wide literature copes with this problem since it is somewhat difficult to separate, without ambiguity, different parts of the signal due to their joint interactions (masking, occlusion, etc.) [3]. In audio processing [4][5][6][7][8][14], the signal is modeled as the superposition of three distinct components: *transients*, that correspond to the attack of the notes or abrupt changes in the sound; *tonal component*, that characterizes the harmonic parts of the signal, and a *stochastic residual*, that is a stationary random signal with smooth power spectrum [6]. A similar hybrid structure can be recognized in shape analysis through a 1-D representation of their contours. It contains information about the local changes of the shape, that can correspond to both natural variations of the curvature or to the spatial discretization of the shape. The former is a transient information, for examples corners, while the latter is a tonal component, since it describes the global orientation of the shape in the space. In Fig. 1 there is an example that shows a simple shape and its normalized differential chain code [9]. This latter is a locally stationary signal with transients in correspondence to abrupt changes in the global orientation. For real world shapes the stochastic characterization of the residual component can be weaker. An example are cavities of historical buildings in cultural heritage. The degradation can be found in the residual component whenever transients and harmonic parts are extracted. The different features of the three components allows to find a proper expansion basis for each of them, i.e. the one yielding a sparse representation. In particular, the tonal component shows

* This paper has been partially supported by the CNR RSTL ‘Multiscale Analysis for Complicated Shapes Recognition’.

significant peaks in the power spectrum, while wavelets are more suitable for transients representation, since their good time localization properties. In audio processing this kind of approach leads to transient and steady state separation (TSS) techniques [4]. Conventional methods first separate the tonal component by selecting significant coefficients in a Fourier basis (Discrete Cosine Transform, Molecular Discrete Cosine Transform, etc.). Then, transients are detected by retaining high amplitude wavelet coefficients of the remaining signal. Another possibility is to define harmonic atoms and use a matching pursuit algorithm for extracting the ones contained in the analysed function [8][10]. Nonetheless, there can be two kinds of problems: i) the signal components cannot be well separated; ii) a thresholding is often employed for detecting significant coefficients. This entails recursive projection algorithm for refining the result. The aim of this work is to exploit the multi-scale characterization of transients for their good detection and separation from the remaining components of the signal. The detection is not limited to give a smooth approximation of the signal (Fourier or wavelet descriptors [9]) but it aims to reconstruct their singular contribution over the original signal. In other words the aim is to have a polygonal representation of the analyzed signal whose corners are in correspondence to true changes in curvature of the signal. We refer to the hybrid signal model used in audio signal, but we try to detect transients directly from the original signal, without a pre-analysis of the tonal component. This is possible thanks to the modeling of transients as isolated singularities in the original signal. They show an increasing decay along scales which is opposite to the tonal component one. This allows their detection at coarser scales while their contribution at finer scales can be predicted using a proper atomic approximation. This latter accounts for transients evolution along scales and corresponds to a piecewise linear representation of the signal. This characterization allows an almost faithful detection and representation of transients and their separation from the tonal component. Some preliminary experimental results will show the potentialities of the method and performances comparable to the TSS techniques.

2 The Proposed Model for Transients Detection

Let the signal $f(t)$ be represented as the superposition of three components:

$$f(t) = f_{ton}(t) + f_{tran}(t) + f_{res}(t), \quad \forall t. \quad (1)$$

The goal of this section will be to extract f_{tran} from f , where f_{ton} is a locally stationary part, i.e. it is the composition of harmonic atoms; f_{tran} is a piecewise smooth signal; f_{res} is a stochastic residual that does not well match with previous classes. The main peculiarity of a wavelet representation is the characterization of singularities through the decay of their coefficients along scale levels [11]. This is known as clustering and persistency property and it is employed in a probabilistic model to discard real transients coefficients. More precisely, a wavelet coefficient can have a transient or a residual state, respectively characterized by a large and small variance distribution. A tree is constructed using the maximum likelihood from coarse to fine scale with the following rule: if the child is a transient, its

parent cannot be a residual [14]. In the following we will use a formal model for describing the evolution of singularities along scales and we will try to use it for transients detection as a direct analysis of the original signal, without a preprocessing for catching the tonal component.

The wavelet theory includes an important result about the characterization of isolated singularities in the time scale. In particular, it is possible to prove that for a signal f of Lipschitz order γ , the decay of the corresponding wavelet coefficient is proportional to $s^{\gamma+1/2}$ [12][1], i.e. $|w(u, s)| \leq As^{\gamma+1/2}$, where $w(u, s)$ is the wavelet transform of f at scale s and time u while A is a constant. In [2], the authors derived the trajectories of modulus maxima corresponding to isolated singularities of order γ_k and they studied their interaction by modeling the coefficients corresponding to a single singularity as waves that travel along scales. These waves are subjected to both a diffusive and source effect till they interfere with other waves. From this point on a transport effect regulates waves interference till they become a single wave. The sourcing effect is weighted by the energy contribution of each wave and it is connected to the order of the singularities. In particular, for the wavelet transform of a piecewise regular signal f with singularities of order γ_k at t_k , it is possible to write the following evolution law in the (u, s) plane:

$$w_s = -\frac{u}{s}w_u + \sum_k \frac{t_k}{s}w_u^{(k)} + \frac{1}{s}w + 2 \sum_k \frac{\gamma_k}{s}w^{(k)}, \quad (2)$$

where w_s and w_u respectively are the partial derivatives of w with respect to s and u , while $w^{(k)}$ is the contribution of the singularity located at t_k . For computational purposes, $w^{(k)}(u, s) = \alpha_k s^{\gamma_k-1} F(t_k, u, s)$, where $F(t_k, u, s)$ is the wavelet transform of a first order singularity located at t_k of a piecewise linear signal and α_k is the corresponding slope — see [2] for details. From the previous equation it is possible to derive the trajectories $u(s)$ of each global maximum of $w(u, s)$ in the time scale plane, i.e.

$$\begin{cases} \dot{u} = -\frac{t_k-u}{s} - \frac{1}{s} \frac{\sum_h d_{kh} w_{uu}^{(h)} + \gamma_h w_u^{(h)}}{w_{uu}} \\ u(1) = t_k \end{cases}$$

where $d_{kh} = t_k - t_h$. If $2C$ is the support for a symmetric and compactly supported wavelet, a singularity is isolated till its cone of influence ($|u - t_k| \leq C s$) does not intersect with the one of another singularities. In other words a singularity can be isolated from scale $s = 1$ till scale $s = \bar{s}$. From \bar{s} on, it interacts with other singularities. From modulus maxima point of view, this movement corresponds to a precise trajectory in the time scale plane: each global maximum is attracted or rejected from the neighboring ones. If it is attracted, it will tend to be confused with the other one, if it is rejected it does not disappear. This kind of interference implies that for a piecewise regular signal the number of global maxima, i.e. the one corresponding to single waves, is constant or decreases along scale levels. On the other hand the tonal component shows a periodic behavior along scale levels, due to its harmonic characterization. It turns out that the

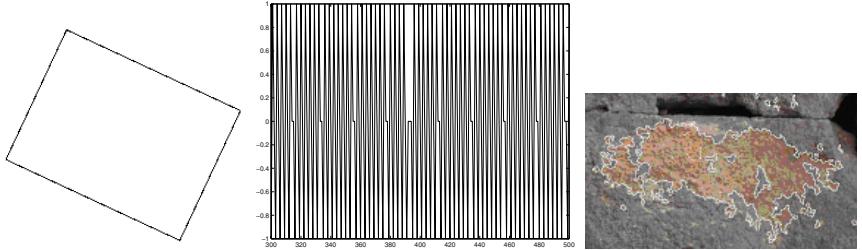


Fig. 1. A rectangular shape with a fixed orientation (*left*). Its corresponding normalized differential chain code (*middle*), that is a piecewise stationary signal. A more complicated shape from degradation of Cultural Heritage (*right*).

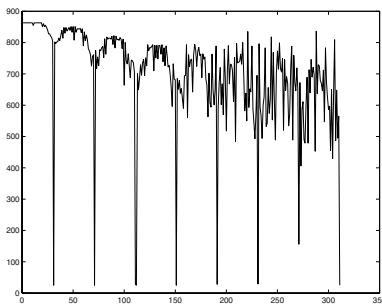


Fig. 2. Number of modulus maxima versus scale of the leftmost shape in Fig. □

numbers of the corresponding modulus maxima could not vary monotonically, as in Fig. □. This observation allows to discard scales where the contribution of the tonal component is not negligible with respect to the transient one. They are the scales such that the monotonic non increasing behavior of the number of modulus maxima is interrupted.

Nonetheless a finite number of scales does not allow the inversion of the transform, while dyadic scales may contain a great contribution of the tonal component. The time scale evolution law of wavelet coefficients in eq. (2) can be exploited to predict details at dyadic scales from the ones selected in the previous step, i.e. the ones having a negligible contribution of the tonal component. In fact, the solution of eq. (2) can be written as $w(u, s) = \sum_k \alpha_k s^{\gamma_k - 1} F(t_k, u, s)$. This form, called atomic approximation, suggests that the wavelet transform of a generic signal at a fixed scale s corresponds to the wavelet transform at that scale of a piecewise linear signal whose slopes are defined by equation $\alpha_{k,s} = \alpha_k s^{\gamma_k - 1}$, as proved by the following proposition.

Proposition 1. *From the atomic approximation at a given scale \bar{s} of a function f , it is possible to build a piecewise linear function $a(t)$ whose wavelet details at the same scale are equal to the ones of a piecewise linear approximation of f .*

Proof: Let us fix the scale level \bar{s} and let us compute the atomic representation of the wavelet transform at that scale, i.e. $\{\alpha_{k,s}, t_k\}_{1 \leq k \leq N}$, hence $w(u, \bar{s}) = \sum_{k=1}^N \alpha_{k,s} F(t_k, u, \bar{s})$, with $\alpha_{k,s} = \alpha_{k,1} s^{\gamma_k - 1}$. Each atom $F(t_k, u, \bar{s})$ equals the wavelet transform at scale \bar{s} of an infinite ramp signal $r_k(t)$ with a singularity in t_k , i.e. $r_k(t) = \begin{cases} 0 & t \leq t_k \\ \alpha_{k,s}(t - t_k) & t > t_k. \end{cases}$ From the linearity of the wavelet transform and indicating with W the wavelet transform operator, it then follows

$$w(u, \bar{s}) = \sum_{k=1}^N \alpha_{k,s} F(t_k, u, \bar{s}) = \sum_{k=1}^N W r_k(t) = W \left(\sum_{k=1}^N r_k(t) \right) = W a(t),$$

where $a(t) = \sum_{k=1}^N \left(\left(\sum_{h=1}^k \alpha_{h,s} \right) (t - t_{h-1}) + \beta_h \right) \chi_{[t_{k-1}, t_k]}$, with $\beta_h = \sum_{h=1}^{k-1} \alpha_{h,s} (t_k - t_h)$. \square

It is worth noticing that the support signal $a(t)$ has the same atomic approximation of the signal f at scale \bar{s} while its low pass residual is different. In that way, from the atomic approximation of the wavelet transform of a signal at a fixed scale \bar{s} , it is possible to derive the details at successive scales by simply performing the wavelet transform of the support signal $a(t)$. While the slopes $\alpha_{k,s}$ can be derived using a matching pursuit like algorithm, as described in [3], the decay exponents γ_k can be extracted by comparing two successive scales at atoms locations. In fact, for two different scales s_1 and s_2 , we have $\alpha_{k,s_1} = \alpha_{k,1} s_1^{\gamma_k - 1}$ and $\alpha_{k,s_2} = \alpha_{k,1} s_2^{\gamma_k - 1}$. Hence, $\alpha_{k,s_2} = \alpha_{k,s_1} \frac{s_2^{\gamma_k - 1}}{s_1^{\gamma_k - 1}}$ and then

$$\gamma_k = 1 + \frac{\log_2 \left(\frac{\alpha_{k,s_2}}{\alpha_{k,s_1}} \right)}{\log_2 \left(\frac{s_2}{s_1} \right)}. \quad (3)$$

It is worth outlining that the atomic representation preserves the correlation between adjacent coefficients of the wavelet decomposition. This property allows the recovering of transients coefficients under threshold and avoids artifacts due to the rough cut off of information in the thresholding based approaches [3].

In order to avoid the introduction of false transients, the coarsest selected scale can provide the intervals where to find transient contribution at finer scales. In fact, as proved by the evolution law, transient contribution becomes more evident at coarser scales. It turns out that they can be selected with a simple thresholding operation. When all the dyadic scales are reconstructed, the transform can be inverted for getting the transient component f_{tran} .

2.1 The Algorithm

1. Perform the continuous wavelet transform (CWT) $w(u, s)$ of f from the scale $s = 1$ to the scale $s = S$ using p as discretization step. Let $J = \lfloor \log_2(S) + 0.5 \rfloor$
2. Compute the low pass component $Af(u, 2^J)$ of f at scale level 2^J
3. Sort the number of modulus maxima of $w(u, s)$ at each scale in increasing order. Let nMAX be the sorted vector

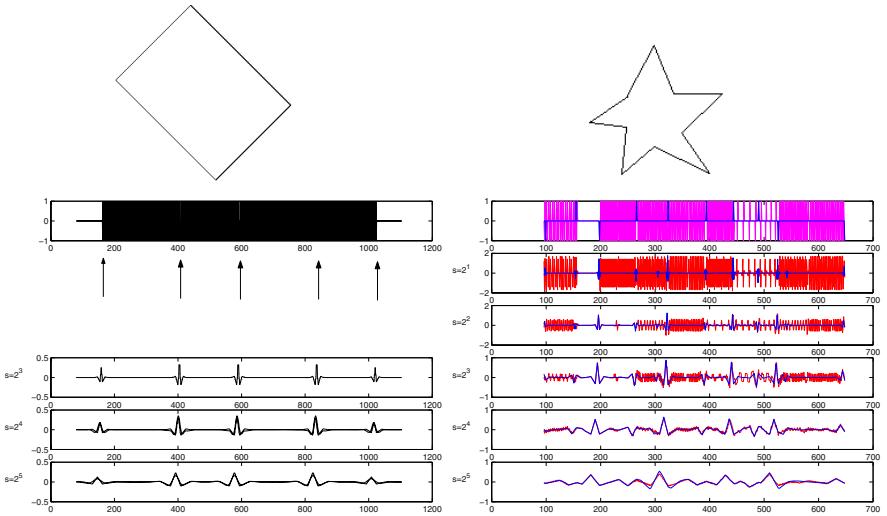


Fig. 3. 1st column) Rectangle oriented at 45°; normalized chain code and transients locations detected using the proposed scheme; reconstructed transient contribution in the wavelet domain. 2nd column) Star shape; the corresponding normalized chain code (magenta) and the estimated f_{trans} (blue) using the proposed scheme; wavelet coefficients of the transient component (blue) using proper basic atoms.

4. Let \hat{s} be the scale level having the number of atoms equal to $nMax(1)$; retain its atoms whose amplitude over-exceeds the value $T = \frac{\|w(u, \hat{s})\|^2}{\hat{s}}$ and select the location of their global maxima $\{t_{k, \hat{s}}\}$
5. For $j = 1, 2, \dots, J$
 - Select the scale \tilde{s} having the least number of maxima among the scale levels that satisfy: $\lfloor \log_2(s) + 0.5 \rfloor = j$
 - Compute the atomic approximation of $w(u, s)$ estimating the atoms slopes α_k using the algorithm in [2], and γ_k using s_j and $s_j + p$ through eq. (3) in the cone of influence of $t_{k, J}$ at the selected scale. Built the corresponding support signal $a(t)$, as in Prop. 1.
 - Compute the undecimated discrete wavelet transform (UDWT) of $a(t)$ at scale level j and let $w(u, 2^j)$ the achieved detail
6. Invert the UDWT using $\{w(u, 2^j)\}_{1 \leq j \leq J}$ and $Af(u, 2^J)$ for getting f_{tran} .

It is worth noticing that step 5 can be relaxed. In fact, from a scale s_j we can predict more than one dyadic scale. In this way it is also possible to drastically reduce the computational effort.

3 Experimental Results and Conclusions

The proposed algorithm has been tested on different test and real world signals. A significant and populated database of shapes (1024 images) and audio signals

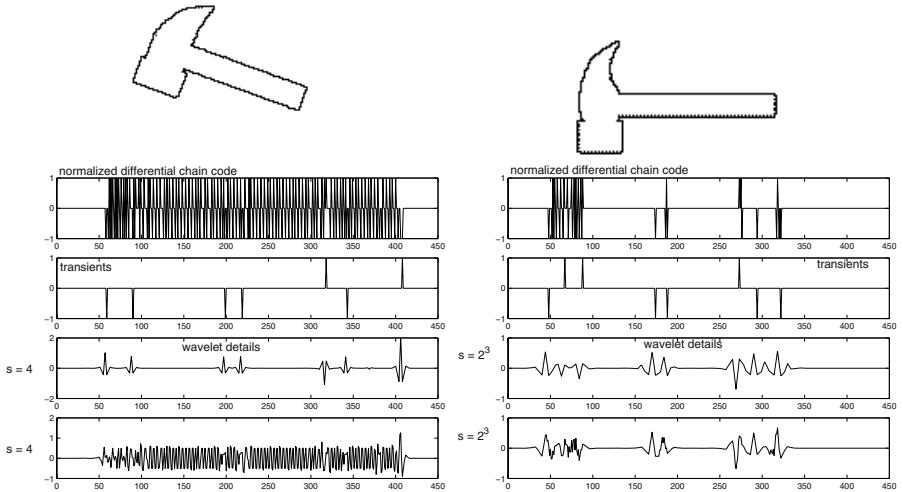


Fig. 4. 1st row) Hammer shape; 2nd row) corresponding normalized differential chain code; 3rd row) transient component estimated using the proposed approach; 4th row) wavelet details at a fixed scale level of the transient component; 5th row) wavelet details at the same scale level of the differential normalized chain code

(300 signals) has been considered while for real shapes we have analysed cavities on the Roman Theatre of Aosta. Achieved results are quite encouraging for both shape analysis and audio processing, since the algorithm selects transient locations almost precisely. In all tests a spline 3/9 biorthogonal wavelet has been used and 5 dyadic scale levels have been considered. The parameter p in step 1 of the algorithm has been set equal to 0.1. The peculiarity of the algorithm is the fact that it tries to reconstruct the transient contribution even at scales where it is dominated by the tonal component (see rightmost figure, in Fig. 3). This allows us to avoid a pre-processing of the signal for detecting the tonal component. In this way it can be used in the processing of the tonal component since it allows to split the signal into distinct stationary pieces. The location of transients is not directly derived from the coarsest scale since they are subjected to transport. Then, it cannot correspond to the actual one in the time domain. For that reason, it is important to recover the contribution of transients at each scale. There are not objective measures for evaluating the final result. For that reason we will provide different examples that are able to show the potential of the method. In Fig. 3 a rectangle and star shapes have been considered. The detected corners are indicated and the wavelet details are compared. It is worth noticing that for differently oriented rectangles, different scale levels have been selected at step 5 of the algorithm, since the different tonal component. For the rectangle oriented at 45°, the selected scales have been $s = 1.1, 3.1, 7.1, 11.2, 31.1$ while for the rectangle oriented at 20°, the scales are $s = 1.1, 3.8, 5.7, 11.3, 31.1$. Fig. 4 depicts two hammer shapes having a different orientation. They provide

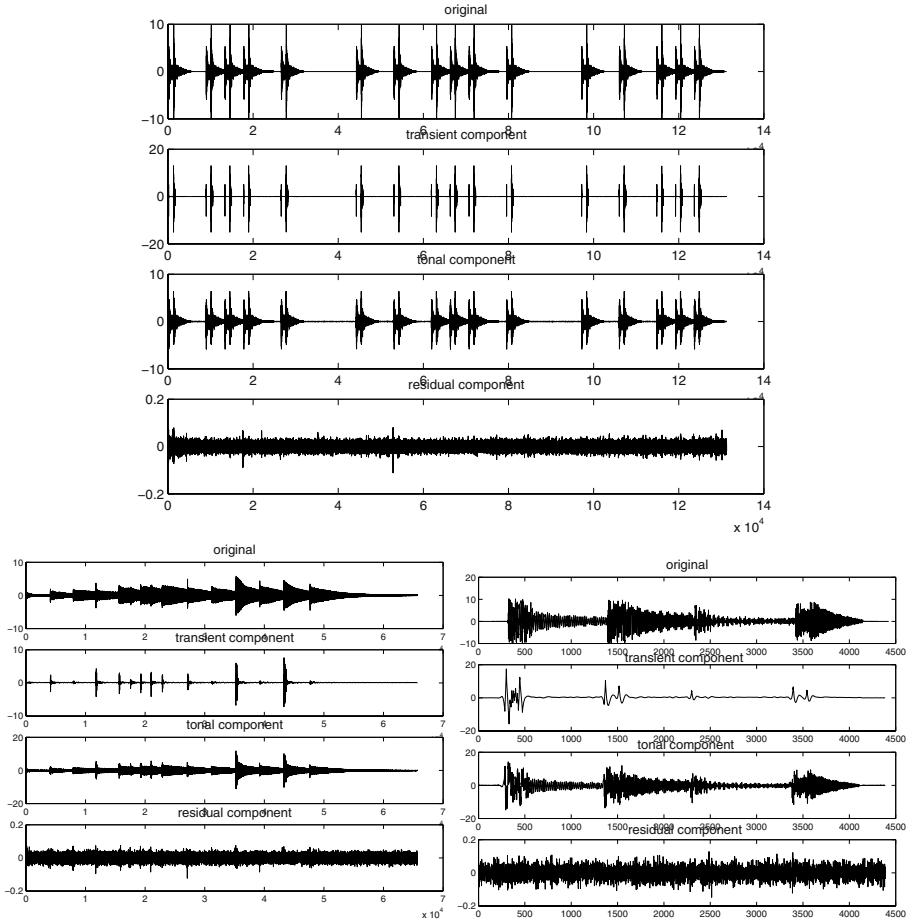


Fig. 5. Castagnette (*top*), Glockenspiel (*bottom left*) and Xilofone (*bottom right*) test audio signals (*top*) and the corresponding three components. f_{tran} has been estimated using the proposed model, while the tonal component has been extracted by thresholding the DCT transform. The residual component is 1/100 of the original signal.

an oscillating normalized differential chain code but their significant corners have been recognized and classified as transient component. Some tests on audio and music signals have been also performed and some results are shown in Fig. 5. Also in this case there is a good detection. To give a measure of the result we used a DCT based techniques for detecting the tonal component and then we measure the energy of the residual. As it can be observed, it is low and it does not show significant contributions in correspondence to transient locations. It is a stochastic signal with a low power spectrum.

With regard to the reconstruction of the transient component, the faithfulness of the representation depends on the adopted basic atom. If for audio signals

we can have a very generic signal, for the normalized chain code, the basic signal is similar to a delta function. Hence, better results can be obtained if the algorithm is applied using the corresponding basic shape in the wavelet domain instead of the piecewise linear one — see right part of Fig. 3. Achieved results are encouraging and model's refinements will be investigated in the future.

References

1. Bello, J.P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., Sandler, M.B.: A Tutorial on Onset Detection in Music Signals. *IEEE Trans. on Speech and Audio Processing* 13(5), 1035–1047 (2005)
2. Bruni, V., Piccoli, B., Vitulano, D.: Wavelet time-scale Dependencies for Signal and Image Compression. In: Pan, Y., Chen, D.-x., Guo, M., Cao, J., Dongarra, J. (eds.) ISPA 2005. LNCS, vol. 3758, pp. 105–110. Springer, Heidelberg (2005)
3. Bruni, V., Vitulano, D.: Wavelet based Signal Denoising via Simple Singularities Approximation. *Signal Processing* 86, 859–876 (2006)
4. Daudet, L.: A Review on Techniques for the Extraction of Transients in Musical Signals. *Computer Music Modeling and Retrieval*, 219–232 (2005)
5. Daudet, L.: Sparse and Structured Decompositions of Signals with the Molecular Matching Pursuit. *IEEE Trans. on Audio, Speech and Language Processing* 14(5) (2006)
6. Daudet, L., Torresani, B.: Hybrid representations for audiophonic signal encoding. *Signal Processing*, Special issue on Image and Video Coding Beyond Standards 82(11), 1595–1617 (2002)
7. Duxbury, C., Davies, M.E., Sandler, M.B.: Separation Of Transient Information In Musical Audio Using Multiresolution Analysis Techniques. In: 4th Int. Workshop on Digital Audio Effects, DAFX 2001, Limerick (2001)
8. Duxbury, C., Chetry, N., Sandler, M., Davies, M.E.: An efficient two-stage implementation of Harmonic Matching Pursuit. In: Proc. of EUSIPCO 2004(2004)
9. Gonzales, R.C., Woods, R.E.: *Digital Image Processing*. Prentice-Hall, Englewood Cliffs (2002)
10. Grinboval, R., Bacry, E.: Harmonic Decomposition of Audio Signals with Matching Pursuit. *IEEE Trans. on Signal Processing* 51(1) (January 2005)
11. Mallat, S., Hwang, W.L.: Singularity Detection and Processing with Wavelet. *IEEE Trans. on Information Theory* 38, 617–643 (1992)
12. Mallat, S.: *A Wavelet Tour of Signal Processing*. Academic Press, London (1999)
13. Molla, S., Torresani, B.: Determining local transientness in audio signals. *IEEE Signal Processing Letters* 11(7), 625–628 (2004)
14. Tantibundhit, C., Boston, J.R., Li, C.C., Durrant, J.D., Shaiman, S., Kovacyk, K., El-Jaraoudi, A.: Speech Enhancement using Transient Speech Components. In: Proc. of IEEE ICASSP 2006 (2006)

SIFT Based Ball Recognition in Soccer Images

Marco Leo, Tiziana D’Orazio, Paolo Spagnolo, Pier Luigi Mazzeo,
and Arcangelo Distante

Institute of Intelligent Systems for Automation

via Amendola 122/D 70126 Bari (Italy)

{leo,dorazio,spagnolo,mazzeo,distante}@ba.issia.cnr.it

Abstract. In this paper a new method for ball recognition in soccer images is proposed. It combines Circular Hough Transform and Scale Invariant Feature Transform to recognize the ball in each acquired frame. The method is invariant to image scale, rotation, affine distortion, noise and changes in illumination. Compared with classical supervised approaches, it is not necessary to build different positive training sets to properly manage the great variance in ball appearances. Moreover, it does not require the construction of negative training sets that, in a context as soccer matches where many no-ball examples can be found, it can be a tedious and long work. The proposed approach has been tested on a number of image sequences acquired during real matches of the Italian Soccer “Serie A” championship. Experimental results demonstrate a satisfactory capability of the proposed approach to recognize the ball.

1 Introduction

During soccer matches a number of doubtful cases occur, especially for detecting the offside or the goal events. An automatic method that detects in each image of the sequence the ball position is the first and the most important step to build a (non invasive) vision based real time decision support tool for the referee committee.

In the last decade different methods to automatically recognize the ball in soccer game have been proposed. They could be conveniently divided in two categories: direct methods and indirect methods. Indirect methods do not search the ball in each frame but they distinguish the ball from other moving objects by means of a priori knowledge about its motion. In [1] a strategy based on no-ball candidate elimination is applied using a coarse-to-fine process. A ‘condensation’ algorithm is utilized in ball tracking and a confidence measure representing the ball region’s reliability is presented to guide possible ball re-detection for continuous tracking. The ball recognition task is achieved in [2] by a trajectory verification procedure based on Kalman filter rather than the low-level feature. Two different procedures run iteratively (trajectory discrimination and extension) to identify the ball trajectory. Soccer ball estimation and tracking using trajectory modeling from multiple image sequences is proposed in [3]. Indirect methods seem to be well suited for video indexing applications but not for real time event detection (as, for example, to solve goal line crossing problem) since they are not able to asses the ball position in each frame.

Direct methods overcome this drawback since they recognize the ball in each frame by using appearance information as shape, size and color. Direct methods classify the pattern images after a suitable pre-processing and they make use of

computational paradigms generally used in many other contexts as face recognition, people detection, character recognition [6-10]. In [14-15] supervised classification techniques based on support vector machine and neural network have been applied in the soccer context on both textured ball and non-textured ball. In [4-5] Wavelet Transform and Independent Component Analysis are used to obtain a suitable representation of the ball to give as input to a neural network.

In this paper a new direct ball recognition approach is proposed. It consists of two steps: first of all the Circle Hough Transform (CHT) [12] is applied on the moving objects in the scene to find the region having the most circular shape inside a given radius; then, the Scale Invariant Feature Transform (SIFT) [11] is applied on the selected region to verify, by analyzing its appearance, if it really contains the ball.

The CHT has been already used for ball recognition purposes [4,5,13] to select regions with circular shape that are candidate to contain the ball. However this step is just preliminary for the actual ball recognition task. Indeed the CHT always produces a maximum value on a region with a circular shape, whatever object is contained inside. Nor the usage of threshold can be considered to select the ball since many objects with circular shape produce high values of the CHT. A further step is necessary to carry out the actual ball recognition process.

The main contribution of this paper is the use of Scale Invariant Feature Transform to validate selected regions. Previous approaches made use of supervised algorithms that required long and tedious learning procedure based on multiple positive training sets (selected manually trying to cover quite all the possible appearances of the ball) to assure acceptable ball recognition performance. Moreover, these methods required a set of negative examples (no-ball examples) to guarantee a more reliable learning phase: unfortunately, the selection of no-ball examples is generally not a trivial task considering that regions with a quite circular shape such as player's socks, pants or shirts, advertising posters, are very common in the soccer images.

The SIFT application allows, in fact, to overcome these drawbacks avoiding the difficulties to build generalized models of the ball and no-ball instances. The proposed approach, taking advantage of the SIFT property of being invariant to image scale, rotation, affine distortion, addition of noise and changes in illumination allows to easily build only one appearance model of the "ball concept" using a small set of ball examples.

In this paper we present a large number of experiments that were carried out on real image sequences acquired during the Italian Soccer "Serie A" championship. We demonstrated that satisfactory ball recognition results can be obtained using only few positive training images.

The rest of the paper is organized as follows: section 2 gives an overview of the proposed method; section 3 describes the experimental setup ; section 4 reports the results of our experiments on real images. Finally, discussion and conclusions are reported in section 5.

2 System Overview

The proposed method consists of two steps (see fig. 1): first of all the Circle Hough Transform (implemented as convolutions on the edge magnitude image) is applied on

the moving regions to select the area that best fits the sought pattern. The region producing the highest peak in the CHT accumulation space is selected as the one having the best circular shape (see [4] for major details).

Then, in the second step a validation procedure is used to verify, by analyzing the appearance, if the selected region really contains the ball. This validation procedure is necessary since the CHT always produces a maximum value in the image, independently from the presence or not of the ball inside the image. The validation procedure proposed in this paper consists in using the SIFT algorithm proposed by Lowe in [11] that extracts distinctive invariant features (the SIFT *keypoints*) by a four stages filtering approach. During an initialization phase a small set of reference ball images is selected and the corresponding SIFT keypoints are stored in a database. This step is carried out just at the beginning during the calibration of the experimental setup and it remains valid for all the experiments.

Then, the SIFT keypoints are evaluated on the region selected by the CHT preprocessing step. These features are compared, by a proper matching procedure (the *keypoint matching procedure* is well described in [11]), to each set of features stored in the database. The final decision about Ball/Non-Ball occurrence in the region is done on the basis of the average number N of correct key-point matches on the stored sets. If N is greater than a fixed threshold Th , experimentally set, the testing image is classified by the system as containing the ball otherwise it is discarded.

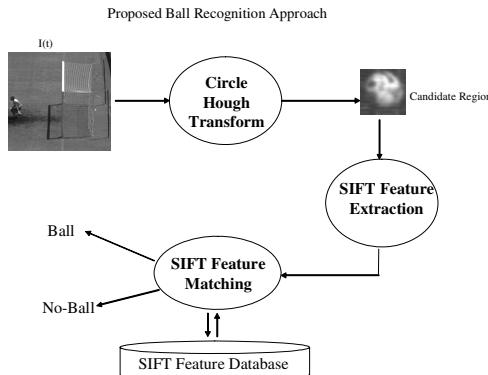


Fig. 1. The Ball recognition system. $I(t)$ is the frame acquired at the time t .

2.1 Scale Invariant Feature Transform (SIFT)

The Scale Invariant Feature Transform is a method for extracting distinctive invariant features from images that can be used to perform reliable matching between different views of an object or scene. The features are invariant to image scale and rotation, and are shown to provide robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination.

The features are highly distinctive, in the sense that a single feature can be correctly matched with high probability against a large database of features from many images.

The algorithm consists of four main steps:

1. Scale-space extrema detection;
2. Keypoints localization;
3. Orientation assignment;
4. Keypoint description.

The first stage identifies locations and scales that can be repeatably assigned under differing views of the same object. Detecting locations that are invariant to scale change of the image can be accomplished by searching for stable features across all possible scales, using a continuous function of scale known as scale space.

Under a variety of reasonable assumptions the only possible scale-space kernel is the Gaussian function. Therefore, the scale space of an image is defined as a function, $L(x, y, \sigma)$, that is produced from the convolution of a variable-scale Gaussian, $G(x, y, \sigma)$, with an input image, $I(x; y)$ i.e.:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)$$

where $*$ is the convolution operation and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}.$$

The keypoints are detected using scale-space extrema in the difference-of-Gaussian function D convolved with the image $I(x; y)$:

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) = L(x, y, k\sigma) - L(x, y, \sigma)$$

where k is the constant multiplicative factor which separates two nearby scales. In order to detect the local maxima and minima of $D(x; y; \sigma)$, each sample point is compared to its eight neighbors in the current image and to its nine neighbors in the scale above and below. It is selected only if it is larger than all of these neighbours or smaller than all of them.

Once a keypoint candidate has been found by comparing a pixel to its neighbours, the next step is to perform a detailed fit to the nearby data for location, scale, and ratio of principal curvatures. This information allows points to be rejected if they have low contrast (and are therefore sensitive to noise) or are poorly localized along an edge. A 3D quadratic function is fitted to the local sample points. The approach starts with the Taylor expansion (up to the quadratic terms) with sample point as the origin

$$D(X) = D + \frac{\partial D^T}{\partial X} X + \frac{1}{2} X^T \frac{\partial^2 D}{\partial X^2} X$$

where D and its derivatives are evaluated at the sample point $X=(x, y, \sigma)^T$. The location of the extremum is obtained taking the derivative with respect to X , and setting it to 0, giving

$$\hat{X} = - \frac{\partial^2 D^{-1}}{\partial X^2} \frac{\partial D}{\partial X}$$

that is a 3x3 linear system, easily solvable.

The function value at the extremum

$$D(\hat{X}) = D + \frac{1}{2} \frac{\partial D^T}{\partial X} \hat{X}$$

is useful for rejecting unstable extrema with low contrast.

At this point the algorithm rejects also keypoints with poorly defined peaks i.e those points having, in the difference-of-Gaussian function a large principal curvature across the edge but a small one in the perpendicular direction.

By assigning a consistent orientation, based on local image properties, the keypoint descriptor can be represented relative to this orientation and therefore achieve invariance to image rotation.

For each image sample, $L(x; y)$, at this scale, the gradient magnitude, $m(x; y)$, and orientation, $\theta(x; y)$, is pre-computed using pixel differences:

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2}$$

$$\theta(x, y) = \text{atan}2((L(x, y+1) - L(x, y-1)) / (L(x+1, y) - L(x-1, y)))$$

An orientation histogram is formed from the gradient orientations of sample points within a region around the keypoint.

Finally a keypoint description is created by first computing the gradient magnitude and orientation at each image sample point in a region around the keypoint location. These sample points are weighted by a Gaussian window and then accumulated into a 8 bins orientation histograms summarizing the contents over 4x4 subregions, with the length of each arrow corresponding to the sum of the gradient magnitudes near that direction within the region. The final descriptor size is $4 \times 4 \times 8 = 128$ and a normalization procedure is applied to the descriptor to allow invariance to illumination change.

3 Experimental Setup

Experiments were performed on real image sequences acquired at the Friuli Stadium in Udine (Italy) during different matches of the “Serie A” Italian Soccer Championship 2006/2007. Images were acquired using a DALSA TM.6740 monochrome cameras able to record up to 200 frames/sec with a resolution of 640x480 pixels. The cameras were placed on the stands of the stadium with theirs optical axis lying on goal-mouth plane (see fig. 2). Different matches in different

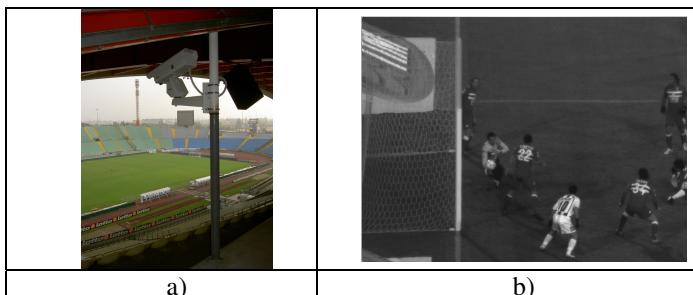


Fig. 2. a) The DALSA TM.6740 camera placed on the stands of the stadium and used to acquire the image sequence during real soccer matches. In figure it is protected by an enclosure. b) an image acquired by the camera during a match.

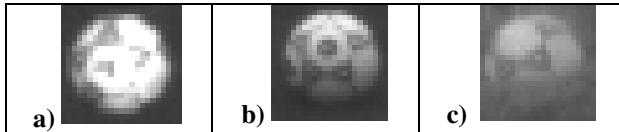


Fig. 3. Three different ball appearances. a) The ball in a sunny day. b) The ball during an evening match. c) The Ball in the goal post. In this case the grid of the goal post is between the camera and the ball.

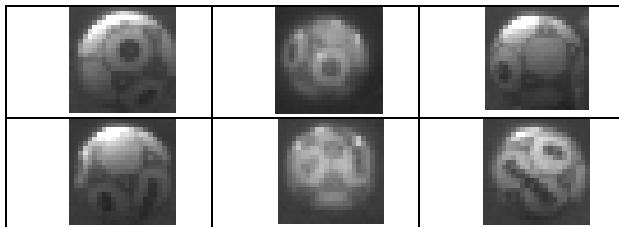


Fig. 4. 6 of the 17 training images used in the experimental phase

lighting conditions (evening matches with artificial lights, afternoon matches in both cloudy and sunny days) were acquired. The acquired images demonstrated the great variance in the appearance of the ball depending on lighting conditions, ball speed, ball position etc. In figure 3 three different ball images are shown.

In the acquired images, the ball radius varied from 9 to 11 pixels (depending on the distance from the camera) so two convolution masks of dimension 23x23 pixels were used to perform Circle Hough Transform and, consequently, a candidate region having size 23x23 pixels was given as input to the validation step based on Scale Invariant Feature Transform.

In the calibration phase, a training set was generated consisting in 17 ball examples acquired during an evening match and chosen to make the system effective also when the ball texture changed under different views. If a uniformly textured ball was used, the training set would be reduced to just one image. The training set had not to take into account possible scale variance of the ball, light condition changes and eventual noise addition. These images were processed to extract the SIFT keypoint features that were saved in a database for further analysis.

In figure 4, some of the 17 training images are reported.

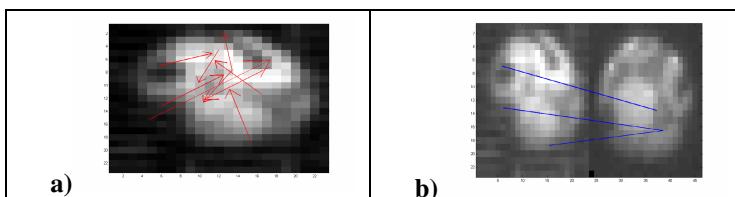


Fig. 5. a) The keypoints localized on a ball image. b) 3 keypoints matching between a test patch and a training patch.

The regions, selected by the CHT step, were classified as a ball instance by the validation step if the average number of key-points matching with the training set keypoints was greater than 3.

In figure 5a) the keypoints detected on a ball image are represented by red arrows; in figure 5b) some keypoints correctly matched between a test image (on the left) and a training image (on the right) are connected by blue lines.

4 Experimental Results

In this paper two sets of experiments were performed: in the first set, the proposed ball recognition approach was evaluated on some sequences acquired during evening matches (with the stadium artificial lights switched on); in the second set of experiments the images were acquired with natural light conditions (both cloudy and sunny conditions).

From the evening matches a set of 3560 images was selected; 1945 of these images contained the ball together with some players and the goal-keeper; the remaining 1615 did not contain the ball but only some players and the goal-keeper.

All these images were firstly processed by the CHT algorithm that, for each of them, selected the region having most circular shape: 1921 of the selected regions really contained the ball, whereas the remaining 1639 did not.

Figure 6 reports some regions selected by CHT algorithm. The two regions on the left contain two ball examples, the two central images contain two examples of partially occluded ball and finally the two images on the right contain two no-ball examples as the shoulder of the goal-keeper and of the shoe of a player. All the regions selected by the CHT algorithm (1921 ball images and 1639 no-ball images), were provided as input to the SIFT based classifier for the validation step that had the main task of separating ball instances from no ball instances.

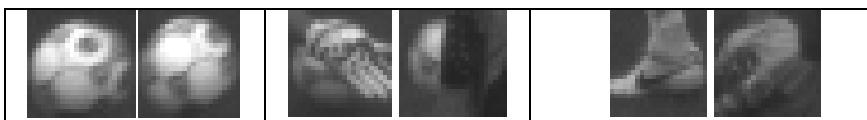


Fig. 6. Some regions selected by CHT algorithm: two examples of whole ball (left column); two examples of partially occluded ball (central column) and two no-ball examples (right column)

In Table 1 the scatter matrix of the results obtained using SIFT based classifier in the first set of experiment is reported. More than 90% of candidate regions containing the ball were correctly validated. At the same time, almost 89% of candidate regions that did not contain the ball were correctly discarded. In figure 7, a case of wrong ball validation is shown. Some keypoints relative to the texture of a player's shoe erroneously matched some keypoints of the ball training images and the proposed approach failed.

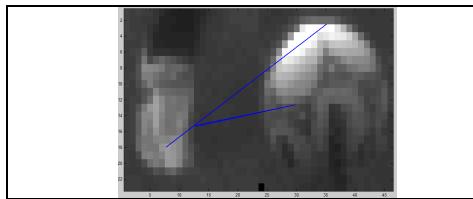


Fig. 7. An example of wrong validation of selected region: the texture of the player's shoe matched some areas of the ball

Table 1. The scatter matrix reporting the performance of the proposed SIFT based classifier in the set of experiments on images acquired during evening matches

		<i>System Results</i>	
		Ball	No Ball
<i>Ground Truth</i>	Ball (1921)	90.3 % (1734/1921)	9.7% (187/1921)
	No-Ball (1639)	10.93% (179/1639)	89.07% (1460/1639)

It should be noted that the CHT algorithm was not able to select the ball in 24 images over the total of 1945 containing the ball. This means that the ball contours were not so clear to produce the maximum value in the CHT accumulation space. In these cases the SIFT classifier is not able to recover the ball, but it can only confirm the absence of the ball in the region erroneously selected by the CHT preprocessing step.

In the second set of experiments, 2147 images were acquired with natural light conditions; 1034 of them contained the ball together with some players and the goal-keeper and the remaining 1113 did not contain the ball but only some players and the goal-keeper.

As in the first experiment these images were processed by the CHT algorithm that for each of them selected the region having the most circular shape: the result was that 1005 of the selected regions really contained the ball, whereas the remaining 1142 did not. Also in this case 29 regions containing the ball were lost by the CHT algorithm that provided maximum values on regions having a more circular appearance than the ball. A careful analysis of these images demonstrated that the ball was not so clear since the images were unfocused (for high speed shot) or the ball was behind the net of the door.

Table 2. The scatter matrix reporting the performance of the proposed SIFT based classifier in the set of experiments on images acquired with natural light conditions

		<i>System Results</i>	
		Ball	No Ball
<i>Ground Truth</i>	Ball (1005)	83.08 % (835/1005)	16.91% (170/1005)
	No-Ball (1142)	10.07% (115/1142)	89.92% (1027/1142)

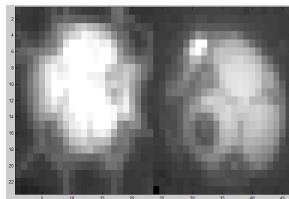


Fig. 8. One example in which the system misrecognized the ball due to reflection effects on the ball surface

In Table 2 the scatter matrix reports the performance results of the SIFT validation step. The presence of self shadows and the saturation of some areas of the ball appearance due to the sun reflection reduced the ball recognition performance with respect to the table 1.

In figure 8 one example of misrecognition of the ball due to the ball saturation effect is shown. The keypoints of the testing image (on the left) did not match with those of the ball in the training set (on the right).

However the general performances remain satisfactory (more than 80% for true positive and true negative detection) also considering that the set of training examples did not contain ball images acquired with natural light conditions. This is, in our opinion, a very pleasant result that makes the proposed approach preferable to those who requires careful selection of training sets.

5 Discussion and Conclusions

In this paper a new method for ball recognition in soccer images is proposed. It combines Circular Hough Transform and Scale Invariant Feature Transform to recognize the ball in each acquired frame. The method is invariant to image scale, rotation, affine distortion, noise, and changes in illumination. It has the main advantage compared with classical supervised approaches, of not requiring different positive training sets to properly manage the great variance in ball appearances. Moreover, it does not require the construction of negative training sets that, in a context as soccer matches where many no-ball examples can be found, it can be a tedious and long work.

Experimental results executed on real images acquired both with natural and artificial lighting conditions, demonstrated the capability of the proposed approach to recognize ball instances. In the reported experiment only one set of 17 ball training images, all acquired during an evening match, was used to performs ball recognition in any lighting condition. This is a very pleasant characteristic for a ball recognition system considering that the ball appearance can greatly change and it is practically impossible to build a generalized ball model.

In conclusion, the proposed approach seems to be a proper trade off between performance, portability and easiness to start up. Future work will be addressed to improve classification performance both using new vision tools able to avoid saturation effect on ball surface in sunny days and introducing more robust strategies to reduce false validations of the ball.

References

- [1] Tong, X.-F., Lu, H.-Q., Liu, Q.-S.: An effective and fast soccer ball detection and tracking method, *Pattern Recognition*, 2004. In: ICPR 2004. Proceedings of the 17th International Conference, August 23-26, 2004, vol. 4, pp. 795–798 (2004)
- [2] Yu, X., Leong, H.W., Xu, C., Tian, Q.: Trajectory-based ball detection and tracking in broadcast soccer video. *IEEE Transactions on Multimedia* 8(6), 1164–1178 (2006)
- [3] Ren, J., Orwell, J., Jones, G.A., Xu, M.: A general framework for 3d soccer ball estimation and tracking. In: ICIP 2004, pp. 1935–1938 (2004)
- [4] D’Orazio, T., Guaragnella, C., Leo, M., Distante, A.: A new algorithm for ball recognition using circle Hough Transform and neural classifier. *Pattern Recognition* 37, 393–408 (2004)
- [5] Leo, M., D’Orazio, T., Distante, A.: Independent Component Analysis for Ball Recognition in Soccer Images. In: Proceeding of the Intelligent Systems and Control ~ISC 2003, Salzburg, Austria (2003)
- [6] Murase, H.: Visual Learning and Recognition of 3-D Objects from Appearance. *International Journal of Computer Vision* 14, 5–24 (1995)
- [7] Papageorgiou, C., Oren, M., Poggio, T.: A general framework for Object Detection. In: Proc. of Intern Conference for Computer Vision (January 1998)
- [8] Rowley, H., Baluja, S., Kanade, T.: Neural Network-Based Face Detection. *IEEE Trans. On Pattern analysis and Machine Intelligence* 20(1), 23–38 (1998)
- [9] Jones, M., Poggio, T.: Mode- based Matching by Linear Combinations of Prototypes. In: Proc. of Image Understanding workshop (1997)
- [10] Mohan, A., Papageorgiou, C., Poggio, T.: Example-based Object Detection in Images by Components. *IEEE Trans. On Pattern analysis and Machine Intelligence* 23(4), 349–361 (2001)
- [11] Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2004)
- [12] Atherton, T.J., Kerbyson, d.J.: Size invariant circle detection. *Image and video Computing* 17, 795–803 (1999)
- [13] Matsumoto, K., Sudo, S., Saito, H., Ozawa, S.: Optimized Camera Viewpoint Determination System for Soccer Game Broadcasting. In: Seo, Y., Choi, S., Kim, H., Hong, K.S. (eds.) Proc. IAPR Workshop on Machine Vision Applications, Tokyo, pp. 115–118 (2000)
- [14] Ancona, N., Cicirelli, G., Branca, A., Distante, A.: Ball recognition in images for detecting goal in football. In: ANNIE 2001, St. Louis, MO (November 2001)
- [15] Ancona, N., Cicirelli, G., Branca, A., Distante, A.: Goal detection in football by using support vector machine for classification. In: International Joint INNS–IEEE Conference on Neural Network, Washington, DC (July 2001)

Evaluation Protocol for Localization Metrics

Application to a Comparative Study

Baptiste Hemery^{1,*}, Hélène Laurent¹, Christophe Rosenberger²,
and Bruno Emile¹

¹ Institut Prisme

ENSI de Bourges - Université d'Orléans

88 boulevard Lahitolle, 18020 Bourges - France

baptiste.hemery@ensi-bourges.fr

² Laboratoire Greyc

ENSICAEN - Université de Caen - CNRS

6 boulevard du Maréchal Juin, 14000 Caen - France

Abstract. Localization metrics permit to quantify the correctness of object detection in an image interpretation result. This paper deals with the definition of a protocol in order to evaluate the behavior of localization metrics. We first define some properties that metrics should verify and create a synthetic database that enables to verify those properties on different metrics. After presenting the tested localization metrics, the results obtained following the proposed protocol are exposed. Finally, some conclusions and perspectives are given.

1 Introduction

Image processing deals with lots of methods from image acquisition (with camera, webcam, satellite...) to image interpretation. Image interpretation consists in extracting information about objects present in an image. Among these information, the automatic localization of an object is a great challenge. As this information is important for many applications, it is required for this localization to be as precise as possible.

In order to evaluate localization algorithms, several research competitions were created such as the Pascal VOC Challenge [1] or the French Robin Project [2]. Given a ground truth, these competitions need a reliable metric to evaluate and compare results obtained by different localization algorithms.

The object localization in an image can be done in many ways: center of the object, bounding box, contour or pixels binary mask. If the localization using the center of the object can be easily evaluated by the Euclidean distance for example, it is not so easy for the other types of localization. By the way, many metrics have been created to evaluate a localization result obtained via a bounding box, a contour or a pixel binary mask. As examples, the Pascal VOC challenge [1] uses a metric based on pixel binary mask, whereas the Robin Project [2]

* The author would like to thank the French Research Ministry for their financial help.

¹ <http://www.pascal-network.org/challenges/VOC/>

² <http://robin.inrialpes.fr/>

created three metrics based on a bounding box. Martin [3] created two metrics based on masks in order to assess segmentation results manually defined done by different persons. Odet [4] also created two metrics, based on contour, in order to evaluate segmentation result. Many other proposals can be found in the literature [5,6,7,8].

This paper aims to define a protocol to evaluate the reliability of a localization metric. This paper is divided in two parts: the first one deals with the definition of required properties for the metrics and the creation of a synthetic database used for the evaluation. The second part presents some tested metrics and give the obtained results. Finally, we conclude and give some perspectives of this study.

2 Evaluation Protocol

A way to evaluate a localization metric is to check if this metric verifies some specific properties. To achieve this goal, we suggest to use the principle of the supervised evaluation of localization, based on a ground truth and a localization result. The metric provides a score corresponding to the coherence between these two images. As we want to verify the properties of metrics, we propose to work in a totally controlled environment using two synthetic images. The first one corresponds to the ground truth and the second one, corresponding to a simulated localization result, is obtained by altering the ground truth (see Fig. 1). As we control the alteration of the ground truth, we can study the evolution of the score given by the localization metric and verify if the metric has the expected properties.

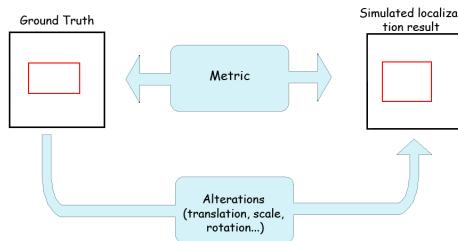


Fig. 1. Protocol Principle

2.1 Properties

We want to verify if a chosen localization metric has good performances regarding to the following properties:

1. Strict Monotony: a metric should penalize the results the more they are altered,
2. Symmetry: a metric should equally penalize two results with the same alteration, but in opposite directions,

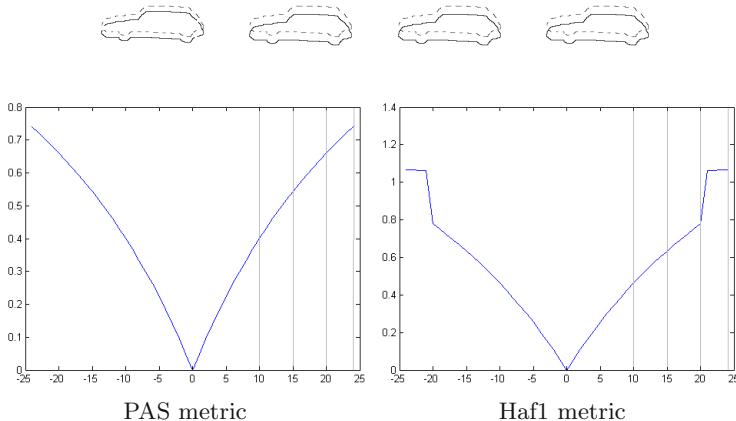


Fig. 2. Examples of localization metrics evaluation: both metrics are symmetric, the first metric is also strictly monotonous and uniformly continuous

3. Uniform Continuity: a metric should not have an important gap between two close results,
4. Topological dependence: a metric result should depend on the size or the shape of the localized object.

We can see on Fig. 2 four cases of alteration of the same ground truth representing a car. This alteration consists in a translation along the vertical axis of 10, 15, 20 and 24 pixels. The two figures below show the evaluation results of similarly altered ground truths by two different metrics. The curves represent the value of the metric in function of the translation along the vertical axis. We can see that both metrics penalize similarly a translation of x pixels and $-x$ pixels. The first metric is monotone and continuous, the result is increasing with the translation and there is no gap. We can see on the second metric that there is a gap when the car is translated of 20 pixels, so the metric is not continuous. Moreover, we can see that after the gap, the metric equally penalizes all alterations, so the metric is not strictly monotonous.

These tested properties have been intuitively chosen and we plan to do a subjective study in order to confirm the importance of these properties for a localization metric evaluation considering a human expect.

2.2 Creation of a Synthetic Database

In order to verify the previously mentioned properties, we need a large amount of images couples corresponding to the ground truth and the simulated localization result. In order to create this database, we considered 16 ground truths that can be seen in Fig. 3. We used 8 ground truths representing a bounding box with different sizes and shapes. We also consider the case where a ground truth is near the border of the image. We also created 8 other ground truths

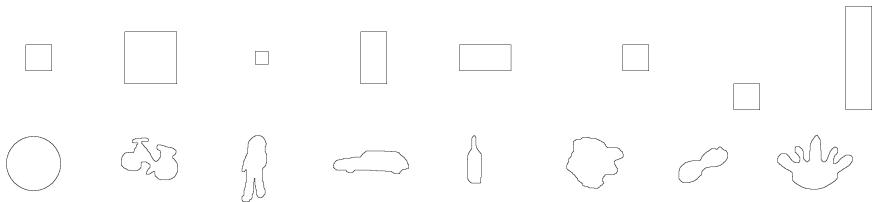


Fig. 3. Ground truths used for the creation of the database

corresponding to real objects that can be used in image interpretation: a bike, a car, a man... Those images are composed of 256 by 256 pixels.

In order to verify if the metrics have the required properties listed before, we used different alterations to create synthetic localization results. We used 4 alterations: translation, scale, rotation and perspective. The translation depends on two parameters: x and y . The parameter x describes a translation along the vertical axis and the y parameter a translation along the horizontal axis. Both parameters evolve between -24 pixels and +24 pixels which leads to 2.400 simulated localization results for one ground truth. The scale alteration depends on two parameters too. The parameter x denotes a scale along the vertical axis and the y parameter a scale along the horizontal axis. Those parameters evolve also between -24 pixels and +24 pixels. A negative value corresponds to a downscaling, whereas a positive value denotes an upscaling. We obtain 2.400 simulated localization results per ground truth. The rotation depends on only one parameter d , corresponding to the angle of rotation in degrees. The parameter d evolves between -90° and $+90^\circ$. This leads to 180 simulated localization results for each ground truth.

The perspective alteration depends on two parameters. The parameter x corresponds to a perspective alteration along the vertical axis and the y parameter a perspective alteration along the horizontal axis. Those parameters evolve also between -24 pixels and +24 pixels. We obtain 2.400 simulated localization results per ground truth.

We can see on Fig. 4 examples of altered ground truth. We finally obtain a total of **118.080** synthetic localization images.



Fig. 4. Examples of altered ground truth

3 Comparative Study

We used this protocol to realize a comparative study of existing localization metrics. Several well known metrics, extracted from the state of the art, were computed on the database allowing to obtain a score for each metric depending of the previously mentioned properties.

3.1 Metrics

There are different types of outputs for a localization algorithm. We present, in the following paragraphs, examples of each type.

The most common one consists in the object localization by a couple of points representing a bounding box. The Robin project [2] aims at evaluating localization and recognition algorithms providing this type of information. In order to achieve this goal, three metrics have been developed to evaluate localization result described by a bounding box:

$$RobLoc(Z_l, Z_{gt}) = \frac{2}{\pi} \arctan(\max(\frac{|x_l - x_{gt}|}{w_{gt}}, \frac{|y_l - y_{gt}|}{h_{gt}})) \quad (1)$$

$$RobCom(Z_l, Z_{gt}) = \frac{|\mathcal{A}_l - \mathcal{A}_{gt}|}{\max(\mathcal{A}_l, \mathcal{A}_{gt})} \quad (2)$$

$$RobCor(Z_l, Z_{gt}) = \frac{2}{\pi} \arctan(|\frac{h_l}{w_l} - \frac{h_{gt}}{w_{gt}}|) \quad (3)$$

where Z_l is the output of the localization algorithm, $\{x_l, y_l\}$ are the coordinates of the center of the bounding box, \mathcal{A}_l is the area covered by the bounding box and $\{h_l, w_l\}$ are the height and width of the bounding box. The variables using $_{gt}$ correspond to the same measures for the ground truth. These three metrics evaluate different characteristics of the localization result: RobLoc evaluates the localization of the center of the bounding box, RobCor quantifies the ratio height/width of the bounding box and RobCom evaluates the size of the bounding box.

There are two other types of localization based on contour or pixels binary mask for representing the localized object. The mask, or region, aspect is, for example, used in the Pascal VOC Challenge [4]. A simple metric is then used to evaluate the localization of an object:

$$PAS(I_{gt}, I_l) = \frac{\text{Card}(I_{gt}^{Re} \cap I_l^{Re})}{\text{Card}(I_{gt}^{Re} \cup I_l^{Re})} \quad (4)$$

with I_l^{Re} corresponds to region pixels of the localized object, $I_{gt}^{Re} \cap I_l^{Re}$ corresponds to object pixels correctly localized and $I_{gt}^{Re} \cup I_l^{Re}$ corresponds to object pixels from the ground truth or from the localized object.

The contour aspect is often used for segmentation evaluation. For example, the Figure of Merit (FOM) proposed by Pratt [5] is an empirical distance between the image with the contour of the localized object I_l and the corresponding ground truth I_{gt} :

$$FOM(I_{gt}, I_l) = \frac{1}{MP} \sum_{k \in I_l^{Cont}} \frac{1}{1 + \alpha * d(k, I_{gt}^{Cont})^2} \quad (5)$$

where I_l^{Cont} are contour pixels of the localized object, MP corresponds to $\text{Max}(\text{Card}(I_{gt}^{Cont}), \text{Card}(I_l^{Cont}))$, α is a constant fixed at $\frac{1}{9}$ and $d(x, I) = \min_{y \in I} d(x, y)$.

Table 1. List of metrics used in the comparative study

Metric	type	reference	Metric	type	reference
RobLoc	Box	[2]	DBad	Contour	[7][9]
RobCom	Box	[2]	ODI	Contour	[4][8]
RobCor	Box	[2]	UDI	Contour	[4][8]
ErrLoc	Contour	[7][8]	PAS	Mask	[4]
ErrSous	Contour	[7][8]	Hen1	Mask	[10]
ErrSur	Contour	[7][8]	Hen2	Mask	[10]
SNR	Contour	[9][11]	Yas1	Mask	[6][8]
RMS	Contour	[9][11]	Yas2	Mask	[6][8]
Lq	Contour	[9][11]	Yas3	Mask	[6][8]
DKu	Contour	[8][12]	Mar1	Mask	[3][8]
DBh	Contour	[8][12]	Mar2	Mask	[3][8]
DJe	Contour	[8][12]	Ham	Mask	[13][8]
DMoy	Contour	[8][14]	Haf1	Mask	[15]
DMoC	Contour	[8][14]	Haf2	Mask	[16]
FOM	Contour	[5][14]	Vin	Mask	[17][18]
DHau	Contour	[7][19]			

For this comparative study, we used 31 localization metrics listed in Tab. 1. Part of those metrics were not created with the specific purpose of localization evaluation, but for the segmentation evaluation or image quality evaluation.

3.2 Experimental Results

The obtained results are presented in Tab. 2. Some metrics depend on an additional parameter, like the distance Lq, DBad, ODI and UDI metrics. We tested these metrics with several values. For each metric, a score is attributed depending upon its properties for each alteration. For example, we attribute a star to one metric if its result to translation is strictly monotonic, a second star if it is symmetric... Therefore, the translation, rotation and perspective alterations maximum score is 5, the scale alteration maximum is 4 because we do not verify if results are symmetric. We also give a final score to each metric which is the sum of scores for each alteration. By the way, the maximum final score is 19.

We can see that metrics used for the Robin project do not have good performance. This comes from the fact that those metrics must be used together in order to correctly evaluate a localization result.

Some metrics based on a contour result: ErrLoc, ErrSous, ErrSur, SNR, RMS, and distances Lq, DBh, DKu, and DJe obtain low scores. This enables to conclude that those metrics should not be used for the evaluation of a localization result. The other metrics based on contour: DMoy, DMoC, FOM, DBad, ODI and UDI metrics obtain quite good results.

Last, all metrics based on pixels binary masks obtain good results. We can note that only the first metric of Martin obtains the maximal score of 19.

Table 2. Synthesis of obtained results

Metric	Type	Translation	Scale	Rotation	Perspective	Final score
RobLoc	Box	*****	—	—	—	5
RobCom	Box	—	***	***	**	8
RobCor	Box	—	***	***	****	10
ErrLoc	Contour	***	**	***	****	12
ErrSous	Contour	***	**	***	****	12
ErrSur	Contour	***	**	***	***	11
SNR	Contour	***	**	**	***	10
RMS	Contour	***	**	****	***	12
Lq,1	Contour	***	**	****	****	13
Lq,3	Contour	***	**	***	***	11
DKu	Contour	***	**	****	****	13
DBh	Contour	***	***	***	**	11
DJe	Contour	***	**	****	****	13
DMoy	Contour	*****	****	****	***	16
DMoC	Contour	*****	****	****	***	16
FOM	Contour	*****	****	*****	****	18
DHau	Contour	***	**	***	***	11
DBad,1	Contour	****	***	*****	***	15
DBad,2	Contour	****	***	****	***	14
DBad,3	Contour	****	**	*****	***	14
ODI,1	Contour	****	**	*****	***	14
ODI,2	Contour	****	**	*****	***	14
UDI,1	Contour	****	**	*****	****	15
UDI,2	Contour	****	**	*****	***	14
PAS	Mask	*****	****	***	*****	18
Hen1	Mask	—	****	—	—	4
Hen2	Mask	*****	****	****	*****	18
Yas1	Mask	*****	***	****	*****	17
Yas2	Mask	*****	***	****	*****	17
Yas3	Mask	*****	***	****	*****	17
Mar1	Mask	*****	****	*****	*****	19
Mar2	Mask	*****	***	****	*****	18
Ham	Mask	****	***	****	*****	16
Haf1	Mask	****	**	****	*****	15
Haf2	Mask	****	***	****	*****	16
Vin	Mask	*****	****	****	*****	18

4 Conclusions

We proposed in this article a protocol that enables to study and compare localization metrics. This protocol allows the comparison of metrics using different representations of a localization result. The comparative study clearly shows that some metrics should not be used for the localization evaluation. Metrics based on pixels binary masks generally give better results than other metrics.

We now plan to realize a subjective evaluation of localization results. This will show the expected properties by a human for localization metrics and will enable to improve our protocol for the evaluation of a localization metrics.

References

1. Everingham, M., Zisserman, A., Williams, C., Van Gool, L., Allan, M., Bishop, C., Chapelle, O., Dalal, N., Deselaers, T., Dorko, G., et al.: The 2005 pascal visual object classes challenge (2005)
2. D'Angelo, E., Herbin, S., Ratiéville, M.: Robin challenge evaluation principles and metrics (2006)
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. In: 8th Int'l Conf. Computer Vision, pp. 416–423 (2001)
4. Odet, C., Belaroussi, B., Benoit-Cattin, H. (eds.): Scalable Discrepancy Measures for Segmentation Evaluation, vol. 1 (2002)
5. Pratt, W., Faugeras, O.D., Gagalowicz, A.: Visual discrimination of stochastic texture fields. *IEEE Transactions on Systems, Man, and Cybernetics* 8(11), 796–804 (1978)
6. Yasnoff, W.A., Mui, J.K., Bacus, J.W.: Error measures for scene segmentation. *Pattern Recognition* 9, 217–231 (1977)
7. Baddeley, A.J.: An error metric for binary images. *Robust Computer Vision*, 59–78 (1992)
8. Chabrier, S.: Contribution à l'évaluation de performances en segmentation d'images. PhD thesis, Université d'Orléans (2005)
9. Wilson, D.L., Baddeley, A.J., Owens, R.A.: A new metric for grey-scale image comparison. *International Journal of Computer Vision* 24(1), 5–17 (1997)
10. Henricsson, O., Baltsavias, E.: 3-d building reconstruction with aruba: A qualitative and quantitative evaluation. In: Automatic Extraction of Man-Made Objects from Aerial and Space Images (II), pp. 65–76 (1997)
11. Coquin, D., Bolon, P., Chehadeh, Y.: Evaluation quantitative d'images filtrées. In: GRETSI 1997, vol. 2, pp. 1351–1354 (1997)
12. Basseville, M.: Distance measures for signal processing and pattern recognition. *Signal Processing* 18(4), 349–369 (1989)
13. Huang, Q., Dom, B.: Quantitative methods of evaluating image segmentation. In: International Conference on Image Processing (ICIP 1995), Washington DC, USA, vol. 3, pp. 53–56 (1995)
14. Peli, T., Malah, D.: A study of edge detection algorithms. *Comp. Graphics Image* (1982)
15. Hafiane, A.: Caractérisation de textures et segmentation pour la recherche d'images par le contenu. PhD thesis, Université de Paris-Sud XI (2005)
16. Hafiane, A., Chabrier, S., Rosenberger, C., Laurent, H.: A new supervised evaluation criterion for region based segmentation methods. In: Blanc-Talon, J., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2007. LNCS, vol. 4678, pp. 439–448. Springer, Heidelberg (2007)
17. Vinet, L.: Segmentation et mise en correspondance de régions de paires d'images stéréoscopiques. PhD thesis, Université de Paris IX Dauphine (1991)
18. Cocquerez, J.P., Philipp, S.: Analyse d'Images: filtrage et segmentation. Masson (1995)
19. Beauchemin, M., Thomson, K.B., Edwards, G.: On the hausdorff distance used for the evaluation of segmentation results. *Canadian Journal of Remote Sensing* 24(1), 3–8 (1998)

Classification of Proteomic Signals by Block Kriging Error Matching

Tuan D. Pham¹, Dominik Beck¹, Miriam Brandl¹, and Xiaobo Zhou²

¹ Bioinformatics Applications Research Center

James Cook University

Townsville, QLD 4811, Australia

² HCNR Center for Bioinformatics

Harvard Medical School

Boston, MA 02115, USA

tuan.pham@jcu.edu.au

Abstract. One of recent advances in biotechnology offers high-throughput mass-spectrometry data for disease detection, prevention, and biomarker discovery. In fact proteomics has recently become an attractive topic of research in biomedicine. Signal processing and pattern classification techniques are inherently essential for analyzing proteomic data. In this paper the estimation method of block kriging is utilized to derive an error matching strategy for classifying proteomic signals with a particular application to the prediction of cardiovascular events using clinical mass spectrometry data. The proposed block kriging based classification technique has been found to be superior to other recently developed methods.

Keywords: Proteomics, mass spectral data, block kriging, signal processing, classification, distortion measures.

1 Introduction

The study of proteomic patterns have recently been utilized for early detection of disease progressions [1][2][3]. Mass spectrometry data has been playing a major role in the discovery of disease pathways and biomarkers for new drug treatment and development [4][5].

Methods for classification of normal and cancerous states using mass spectrometry (MS) data have been recently developed. Petricoin *et al.* [2] applied cluster analysis and genetic algorithms to detect early stage ovarian cancer using proteomic spectra. Ball *et al.* [6] applied integrated approach based on neural networks to study SELDI-MS data for classification of human tumors and identification of biomarkers. Lilien *et al.* [7] applied principal component analysis and a linear discriminant function to classify ovarian and prostate cancers. Sorace and Zhan [8] used mass spectrometry serum profiles to detect early ovarian cancer. Wu *et al.* [9] compared the performance of several methods for the classification of mass spectrometry data. Tibshirani *et al.* [10] proposed a probabilistic approach for sample classification from protein mass spectrometry data. Morris

et al. [11] applied wavelet transforms and peak detection for feature extraction of MS data. Yu *et al.* [12] developed a method for dimensionality reduction for high-throughput MS data. Levner [13] used feature selection methods and then applied the nearest centroid technique to classify MS-based ovarian and prostate cancer datasets.

Given the promising integration of several classification methods and mass spectrometry data in high-throughput proteomics [14], this new biotechnology still encounters several challenges in order to become a mature platform for clinical diagnostics and protein-based biomarker profiling. One of a major concerns is the finding of an effective computational approach for the analysis of this type of high-throughput data. In this paper we discuss for the first time the implementation of blocking kriging technique to determine the long-range spatial error variances of mass spectrometry data that can be used as a basis for signal matching and classification.

2 Error Matching by Block Kriging

Kriging techniques [15][16] estimate the unknown value at a particular location as the linear combination of the known values at nearby locations:

$$\hat{s}(n) = \sum_{k=1}^p w_k s(n_k) \quad (1)$$

where w_k , $k = 1, \dots, p$ are the kriging weights, and $s(n_k)$, $k = 1, \dots, p$, are the known data values at locations n_k .

The central idea to this estimation is to determine a set of optimal kriging weights. Using the method of block kriging, these optimal weights can be obtained as

$$\mathbf{Cw} = \mathbf{b} \quad (2)$$

where \mathbf{C} is the square and symmetrical matrix that represents the spatial covariances between the known signals, \mathbf{w} is the vector of kriging weights and a Lagrange multiplier μ , and \mathbf{b} is the vector that represents the average spatial covariances between a particular sample location and all the points within a domain A :

$$\mathbf{C} = \begin{bmatrix} C_{11} & \cdots & C_{1n} & 1 \\ \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdots & \cdot & \cdot \\ C_{n1} & \cdots & C_{nn} & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix}$$

$$\mathbf{w} = [w_1 \cdots w_n \mu]^T$$

$$\mathbf{b} = [C_{1A} \cdots C_{nA} 1]^T$$

The sample spatial covariance used for the kriging estimator can be calculated as

$$C_{ij} = \frac{1}{N(h)} \sum_{(i,j)|h_{ij}=h} s(j) - \left(\frac{1}{n} \sum_{k=1}^n s(k) \right)^2 \quad (3)$$

in which the sample spatial covariance is a function of the lag distance h , $N(h)$ is the number of pairs that $s(i)$ and $s(j)$ are separated by h , and n is the total number of data points.

The average spatial covariances between a sample location and all the points within A is defined as

$$C_{iA} = \frac{1}{N} \sum_{j|j \in A}^N C_{ij} \quad (4)$$

Thus the vector of the spatial predictor coefficients can be obtained by solving

$$\mathbf{w} = \mathbf{C}^{-1} \mathbf{b} \quad (5)$$

The block kriging error variance is given by

$$\sigma_{BK}^2 = C_{AA} - \mathbf{w}^T \mathbf{b} \quad (6)$$

where

$$C_{AA} = \frac{1}{NM} \sum_{i|i \in A}^M \sum_{j|j \in A}^N C_{ij} \quad (7)$$

In terms of the semi-variogram, the block kriging error variance is given by

$$\sigma_{BK}^2 = \mathbf{w}^T \mathbf{b} \quad (8)$$

where all the covariance terms involving in \mathbf{C} and \mathbf{b} expressed in (2) are replaced with the semi-variogram values.

The semi-variogram is a function which expresses the spatial relationship of a regionalized variable [17]. In probabilistic notation, the variogram, $2\gamma(h)$, is defined as the expected value:

$$2\gamma(h) = E\{[s(i) - s(j)]^2\}, h_{ij} = h \quad (9)$$

where h is a lag distance that separates $s(i)$ and $s(j)$.

The semi-variogram is half of the variogram, that is, $\gamma(h)$. The experimental semi-variogram for lag distance h is defined as the average squared difference of values separated by h :

$$\gamma(h) = \frac{1}{2N(h)} \sum_{(i,j)|h_{ij}=h} [s(i) - s(j)]^2 \quad (10)$$

where $N(h)$ is the number of pairs for lag h .

We now consider \mathbf{x} , \mathbf{y} , and \mathbf{z} as the vectors defined on a vector space V . A metric or distance d on V is defined as a real-valued function on the Cartesian product $V \times V$ if it has the properties of positive definiteness, symmetry, and triangle inequality. If a measure of dissimilarity satisfies only the property of positive definiteness, it is referred to as a distortion measure which is considered very common for the vectorized representations of signal spectra [18]. In general, to calculate a distortion measure between two vectors \mathbf{x} and \mathbf{y} , denoted as $D(\mathbf{x}, \mathbf{y})$, is to calculate a cost of reproducing any input vector \mathbf{x} as a reproduction of vector \mathbf{y} . Given such a distortion measure, the mismatch between two signals can be quantified by an average distortion between the input and the final reproduction. Intuitively, a match of the two patterns is good if the average distortion is small.

A very useful distortion measure that is derived from the above mathematical basis is the likelihood ratio distortion between the two templates presented in the form of two vectors of predictor coefficients \mathbf{w} , and \mathbf{w}' which are used to model signal s . The likelihood-ratio distortion measure, denoted by D_{LR} , is defined as [18]

$$D_{LR}(\mathbf{w}, \mathbf{w}') = \frac{\mathbf{w}'^T \mathbf{R}_s \mathbf{w}'}{\mathbf{w}^T \mathbf{R}_s \mathbf{w}} - 1 \quad (11)$$

where \mathbf{R}_s is the autocorrelation matrix of sequence s associated with its prediction coefficient vector \mathbf{w} , and \mathbf{w}' is the prediction coefficient vector of signal s' . For a perfect match between the two templates, the errors are identical and (11) yields a zero distortion. For a mismatch, the residual resulting from the prediction analysis is large and the distortion defined in (11) therefore becomes large.

Based on the same principle derived for the likelihood ratio distortion, the block kriging distortion measure, denoted as D_{BK} , can be defined as

$$D_{BK}(\mathbf{w}, \mathbf{w}') = \frac{\mathbf{w}'^T \mathbf{b}}{\mathbf{w}^T \mathbf{b}} - 1 \quad (12)$$

where \mathbf{w} is defined in (2) which is the kriging vector of signal s , \mathbf{b} is the vector, in terms of the semi-variogram values, defined in (2) associated with s , and \mathbf{w}' is the kriging vector of signal s' .

If the input (unknown) MS signal s_m is analyzed by the prediction analysis which results in a set of prediction coefficients, then the spectral distortion between an unknown sample s_m and a particular known class i can be determined using the minimum rule as follows.

$$D_{min}(\mathbf{x}_m, \mathbf{c}^i) = \min_j D(\mathbf{x}_m, \mathbf{c}_j^i) \quad (13)$$

where D is a spectral distortion measure, \mathbf{x}_m is the prediction vector of s_m , \mathbf{c}_j^i is the prediction vector of the j sample that belongs to class i .

Using a simple decision logic, the unknown signal s_m is assigned to class i^* if the minimum distortion measure of its prediction vector \mathbf{x}_m and the corresponding prediction vector \mathbf{c}^i is minimum, that is

$$s_m \rightarrow i^*, \quad i^* = \arg \min_i D_{min}(\mathbf{x}_m, \mathbf{c}^i) \quad (14)$$

3 Experimental Results

We used high-throughput, low-resolution SELDI MS (www.ciphergen.com) to acquire the protein profiles from patients and controls. The protein profiles were acquired from 2 kDa to 200kDa. The design of the experiment originally described in [19] involves the datasets for the control and MACE group.

For the control group, the dataset consists of sixty patients who presented in emergency room with chest pain and the patients' troponin T test was consistently negative. These patients lived in the next 5 years without any major cardiac events or death. The total 166 plasma samples, 24 reference samples and 6 blanks were fractionated into 6 fractions using two 96-well plates containing anion exchange resin (Ciphergen, CA).

For the MACE group, the dataset was designed to comprise 60 patients who presented in emergency room with chest pain but the patients' troponin T test was negative. However, the patients in this group had either a heart attack, died or needed revascularization in the subsequent 6 months. The blood samples used in this study were same as those used in [20]. Most new MPO data measured with FDA approved CardioMPO kit for these two groups are available – MPO levels for 56 (out of 60) patients in control group and 55 (out of 60) patients in MACE group are available. Statistical analysis shows that MPO alone can distinguish MACE from control with accuracy of better than 60%.

For the SELDI mass spectra, the coverage of proteins in SELDI protein profiles was increased by that the blood samples were fractionated with HyperD Q (strong ion exchange) into 6 fractions. The protein profiles of fractions were acquired with two SELDI Chips: IMAC and CM10. There are a few different SELDI chips with different protein binding properties. Generally speaking, the more types of the SELDI chips are used, the more proteins are likely to be detected. However, due to the high concentration dynamic range of the proteins in

Table 1. Average MACE prediction rates by different methods

Method	Average accuracy (%)
MPO value	55.25
T-test	62.23
Standard genetic algorithm	69.05
Sequential forward floating search	71.92
Improved genetic algorithm	75.16
Block kriging distortion measure	93.15

human blood, the total number of proteins to be detected by the protocol we are using is very limited. We estimate that the number of the proteins we are able to detect is about one-thousand, while the total protein number in human blood is estimated to be tens of thousands. For example, MPO can be accurately measured with immunoassay (CardioMPO) but could not be detected with SELDI MS. The MS data for each sample in each fraction was acquired in duplicate, so 120 samples (60 controls and 60 MACEs) in each fraction in one type of SELDI chip have 240 spectra. There are two types of SELDI chips: IMAC and CM10.

To emphasize our study on rigorous prediction using SELDI mass spectra, we used only two fractions to carry out the experiment. Because of the short length of the samples, we concatenated the corresponding samples of the two fractions for the extraction of the prediction coefficients. We have recently applied the statistical and geostatistical prediction models, and the prediction based classification rule for extracting the MS features and classifying control and MACE samples, respectively [21]. The pattern matching using block kriging error matching defined in [12] was carried out for the same dataset in this study.

In previous work [21], we performed the leave-one-out validation and obtained the average classification rate of 83.34% using the statistical distortion measure; whereas for the ordinary kriging distortion measure, the average classification result was 97.10%. However, these classification results were based on the use of whole MS sequences and not based on the MS peak values. In another previous work [19], we used the MPO value, five selected biomarkers by *T*-test, five selected biomarkers by the sequential forward floating search (SFFS), five selected biomarkers by standard genetic algorithm (GA), and five selected biomarkers by an improved genetic algorithm (IGA) to carry out the prediction. Using the same dataset and test design, the block kriging gave the average classification result as 93.15%. The average validation results of different methods are presented in Table I which shows the superior performance of the block kriging approach over other relevant models.

4 Conclusion

It has been predicted that the advancement of proteomics pattern diagnostics might represent a revolution in the field of molecular medicine, because this technology has the potential of developing a new model for early disease detection [3][22][23]. Given that the research into clinical proteomic pattern diagnostics is still in its infancy because the results have not been validated in large trials, and effective computational methods have not been well-explored; recent research outcomes have illustrated the role of MS-based proteomics as an indispensable tool for molecular and cellular biology and for the emerging field of systems biology [5].

Early disease detection using MS data is a challenging task up to date. This task requires the combination of the contrast fields of knowledge of modern biology and computational methodology. We have presented in this paper the novel applications of the theories of linear prediction in time and space domains for

extracting the effective features of mass spectrometry data that can be useful for the classification of MS spectra. The initial results using the cardiovascular SELDI-MS datasets have shown the potential application of the proposed techniques for predicting patient's major adverse cardiac risk that can be helpful for the development of the diagnostic kit for treatment of patients during their initial emergency admission [24]. The use of the prediction coefficients can be extended to other spectral coefficients derived from the prediction principle, and the distortion measures can be used as similarity measures for other classification techniques.

Research into MS-based disease detection has recently attracted the attention of researchers from various disciplines. In particular, it offers tremendous potentials for the discovery of novel biomarkers and the development of personalized medicine [25] – a new concept that major diseases have a genetic component; therefore the understanding of cellular processes at the molecular level will enable scientists and physicians to predict the relative risk and potential therapy for such conditions on a person-to-person basis.

Acknowledgments

This research was supported under the Australian Research Council's Discovery Projects funding scheme (project number DP0877414) to Tuan Pham.

References

1. Sauter, E., et al.: Proteomic analysis of nipple aspirate fluid to detect biologic markers of breast cancer. *Br. J. Cancer* 86, 1440–1443 (2002)
2. Petricoin, E.F., et al.: Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 359, 572–577 (2002)
3. Conrads, T.P., Zhou, M., Petricoin III, E.F., Liotta, L., Veenstra, T.D.: Cancer diagnosis using proteomic patterns. *Expert Rev. Mol. Diagn.* 3, 411–420 (2003)
4. Griffin, T., Goodlett, T., Aebersold, R.: Advances in proteomic analysis by mass spectrometry. *Curr. Opin. Biotechnol.* 12, 607–612 (2002)
5. Aebersold, R., Mann, M.: Mass spectrometry-based proteomics. *Nature* 422, 198–207 (2003)
6. Ball, G., Mian, S., Holding, F., Allibone, R.O., Lowe, J., Ali, S., Li, G., McCardle, S., Ellis, I.O., Creaser, C., Rees, R.C.: An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. *Bioinformatics* 18, 395–404 (2002)
7. Lilien, R.H., Farid, H., Donald, B.R.: Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. *J. Computational Biology* 10, 925–946 (2003)
8. Sorace, J.M., Zhan, M.: A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* 4, 24 (2003)
9. Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., Zhao, H.: Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* 19, 1636–1643 (2003)

10. Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A., Le Sample, Q.-T.: classification from protein mass spectrometry, by peak probability contrasts. *Bioinformatics* 20, 3034–3044 (2004)
11. Morris, J.S., Coombes, K.R., Koomen, J., Baggerly, K.A., Kobayashi, R.: Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics* 21, 1764–1775 (2005)
12. Yu, J.S., Ongarello, S., Fiedler, R., Chen, X.W., Toffolo, G., Cobelli, C., Trajanoski, C.Z.: Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. *Bioinformatics* 21, 2200–2209 (2005)
13. Levner, I.: Feature selection and nearest centroid classification for protein mass spectrometry. *BMC Bioinformatics* 6, 68 (2005)
14. Shin, H., Markey, M.K.: A machine learning perspective on the development of clinical decision support systems utilizing mass spectra of blood samples. *J. Biomedical Informatics* 39, 227–248 (2006)
15. Matheron, G.: The theory of regionalized variables and its applications. Paris School of Mines Publication, Paris (1971)
16. Isaaks, E.H., Srivastava, R.M.: An Introduction to Applied Geostatistics. Oxford University Press, New York (1989)
17. Deutsch, C.V.: Geostatistical Reservoir Modeling. Oxford University Press, New York (2002)
18. Rabiner, L., Juang, B.-H.: Fundamentals of Speech Recognition. Prentice Hall, New Jersey (1993)
19. Zhou, X., Wang, H., Wang, J., Hoehn, G., Azok, J., Brennan, M.L., Hazen, S.L., Li, K., Wong, S.T.C.: Biomarker discovery for risk stratification of cardiovascular events using an improved genetic algorithm. In: Proc. IEEE/NLM Int. Symposium on Life Science and Multimodality, pp. 42–44 (2006)
20. Brennan, M.-L., Penn, Van Lente, M.S., Nambi, V., Shishehbor, M.H., Aviles, R.J., Goormastic, M., Pepoy, M.L., McErlean, E.S., Topol, E.J., Nissen, S.E., Hazen, S.L.: Prognostic value of myeloperoxidase in patients with chest pain. *The New England Journal of Medicine* 348, 1595–1604 (2003)
21. Pham, T.D., Wang, H., Zhou, X., Beck, D., Brandl, M., Hoehn, G., Azok, J., Brennan, M.-L., Hazen, S.L., Li, K., Wong, S.T.C.: Computational prediction models for early detection of risk of cardiovascular events using mass spectrometry data. *IEEE Transactions on Information Technology in Biomedicine* (in print)
22. Petricoin, E.F., Liotta, L.A.: Mass spectrometry-based diagnostics: The upcoming revolution in disease detection. *Clinical Chemistry* 49, 533–534 (2003)
23. Wulfkuhle, J.D., Liotta, L.A., Petricoin, E.F.: Proteomic applications for the early detection of cancer. *Nature* 428, 267–275 (2003)
24. Goodacre, S., Locker, T., Arnold, J., Angelini, K., Morris, F.: Which diagnostic tests are most useful in a chest pain unit protocol? *BMC Emergency Medicine* 5, 6 (2005)
25. Ginsburg, G.S., McCarthy, J.J.: Personalized medicine: revolutionizing drug discovery and patient care. *Trends Biotechnol.* 19, 491–496 (2001)

Grip-Pattern Recognition in Smart Gun Based on Likelihood-Ratio Classifier and Support Vector Machine

Xiaoxin Shang, and Raymond N.J. Veldhuis

Signals and Systems Group, Electrical Engineering, University of Twente
P.O. Box 217, 7500 AE Enschede, the Netherlands
x.shang@utwente.nl

Abstract. In the biometric verification system of a smart gun, the rightful user of a gun is recognized based on grip-pattern recognition. It was found that the verification performance of this system degrades strongly when the data for training and testing have been recorded in different sessions with a time lapse. This is due to the variations between the probe image and the gallery image of a subject. In this work the grip-pattern verification has been implemented based on both classifiers of the likelihood-ratio classifier and the support vector machine. It has been shown that the support vector machine gives much better results than the likelihood-ratio classifier if there are considerable variations between data for training and testing. However, once the variations are reduced by certain techniques and thus the data are better modelled during the training process, the support vector machine tends to lose its superiority.

1 Introduction

We develop a prototype recognition system as part of a smart gun, where the hand-grip pattern recognition ensures that the gun can only be fired by the rightful user. This system is intended to be used by the police, since carrying a gun in public brings considerable risks. In the US, for example, vital statistics show that about 8% of the law-enforcement officers killed in a shooting incident were shot by their own weapons [1].

Figure 1 shows both the prototype of the smart gun and an example of the grip-pattern image. One can see from the right-side figure the pressure pattern of the thumb in the upper-left corner of the image, and those of fingers in the remaining part. Note that only three fingers are present, because the index finger is near the trigger of the gun. We collected the grip-pattern data from a group of police officers in three sessions with a time lapses in between [2]. The data were processed for verification based on a Likelihood-Ratio Classifier (LRC) described in [3]. Initial experimental results indicate that if data for training and testing come from the same collection session, the verification results are fairly good, with an equal-error rate (EER) below 1%; otherwise the results are unsatisfactory, i.e., about 15% EER on average. Since in practice there will always be a time interval between data enrollment and verification, the across-session results are more relevant and, therefore need to be improved.

Having analyzed the data collected all sessions, we found that the data of one subject collected across sessions vary greatly, even though the grip-pattern images of this subject from one same session look fairly similar [2]. There are mainly two types of

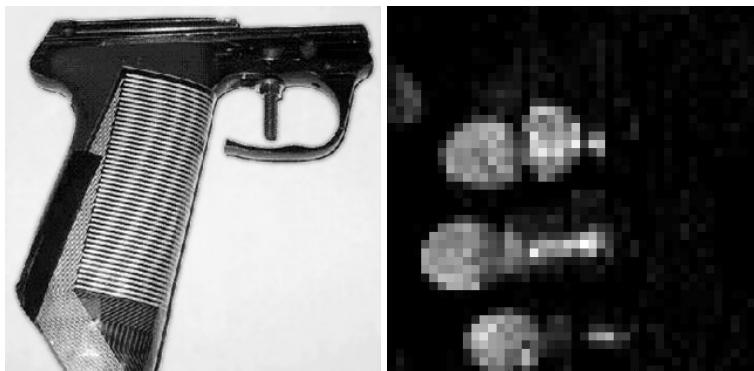


Fig. 1. Left: prototype of the smart gun. Right: an example of grip-pattern image.

across-session variations. First, a variation of pressure distributions occurs between grip patterns from a subject collected in different sessions. A second type of variation results from the hand shift of a subject across sessions [2]. Figure 2 shows two images collected from one subject in two different sessions, respectively. One can see that these two images have quite different pressure distributions. Besides, the hand-grip pattern in the image on the right side is located higher, than that in the image on the left side. Further research showed that these variations are the main reason for the unsatisfactory across-session verification results [2].

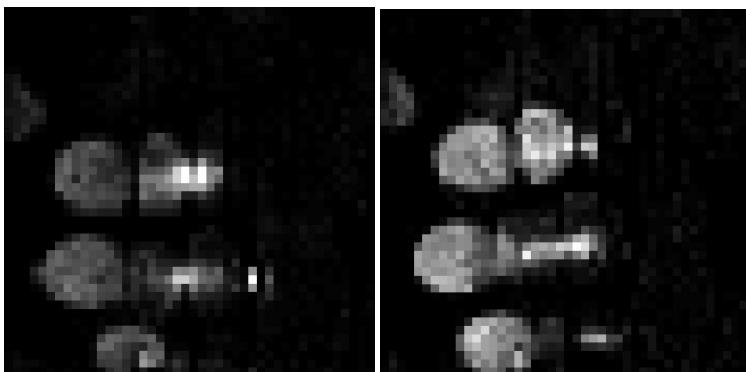


Fig. 2. Grip-pattern images of a subject in different collection sessions

Based on the characteristics of the grip-pattern images mentioned above, the verification results can be improved by reducing the across-session variations of data. In earlier work we applied three methods, each of which effectively improved the verification results, respectively. Firstly, we used template-matching registration (TMR) to reduce the across-session variation due to the hand shift [4][5]. By doing this the EER was reduced

to about 13% from about 15%. The second technique that we applied was the double-trained model (DTM), where the data from two out of three collection sessions were combined for training, and those of the remaining session were used for testing. With DTM, the across-session variations of data were much better modelled in the training procedure, compared to the case where only one collection session of data were used for training. The verification results proved to be greatly improved by DTM, with the EER reduced from 15% to about 8% on average. Thirdly, we applied an image preprocessing approach, Local Absolute Binary Patterns (LABP), prior to classification [6]. This technique can reduce the across-session variation of hand-pressure distribution. Specifically, with respect to a certain pixel in an image, the LABP processing quantifies how its neighboring pixels fluctuate. It was found that the application of LABP improved the verification results significantly, with the EER reduced from 15% to about 9% on average. Finally, the verification results were improved greatly when all these three methods were applied together, with an average EER of about 3% approximately.

Note that all the verification results given above are based on LRC, which requires estimation of the probability density function (PDF) of the data [3]. Therefore, if there exist large variations between data for training and testing, the verification results will be greatly degraded. To further improve the verification results and also to set a reference for evaluation of the results obtained so far, we decided to implement another classifier which is more capable to cope with the across-session variations of data. With these motives, we chose the Support Vector Machine (SVM). As a contrast to LRC, SVM does not estimate the data's distribution. Instead, it tries to maximize the margin between different classes. Therefore, it is expected to be more robust to across-session data variations than the PDF-based classifiers in a many cases.

This paper presents and compares the verification results by using SVM and LRC. The remainder of this paper is organized as follows: Section 2 briefly describes the verification algorithms based on LRC and SVM, respectively. Subsequently, Section 3 presents and discusses the experimental results. Finally, conclusions are given in Section 4.

2 Verification Algorithms

2.1 Likelihood-Ratio Classifier

In classification by the LRC, it is assumed that the data is Gaussian [7], [8]. The pixel values of a grip-pattern image are arranged into a (in this case $44 \times 44 = 1936$ -dimensional) column vector \mathbf{x} . The feature vector \mathbf{x} is normalized, i.e. $\|\mathbf{x}\|_2 = 1$, prior to classification. A measured image originates either from a genuine user, or from an impostor. The grip-pattern data of a certain subject is characterized by a mean vector $\boldsymbol{\mu}_W$ and a covariance matrix $\boldsymbol{\Sigma}_W$, where the subscript W denotes 'Within-class'; while the impostor data is characterized by $\boldsymbol{\mu}_T$ and $\boldsymbol{\Sigma}_T$, where the subscript T denotes 'Total'. The matching score of a measurement \mathbf{x} with respect to this subject is derived from the log-likelihood ratio. It is computed by

$$\begin{aligned} S(\mathbf{x}) = & -(\mathbf{x} - \boldsymbol{\mu}_W)^T \boldsymbol{\Sigma}_W^{-1} (\mathbf{x} - \boldsymbol{\mu}_W) \\ & + (\mathbf{x} - \boldsymbol{\mu}_T)^T \boldsymbol{\Sigma}_T^{-1} (\mathbf{x} - \boldsymbol{\mu}_T). \end{aligned} \quad (1)$$

If $S(\mathbf{x})$ is above a preset threshold, the measurement is accepted as being from the genuine user. Otherwise it is rejected [3]. The threshold determines the false reject rate (FRR) and the false acceptance rate (FAR) of verification.

In practice the mean vectors and covariance matrices are unknown, and need to be estimated from a set of training data. In our case, the number of training samples from each subject should be much greater than 1936. Otherwise, the algorithm would become overtrained [3]. However, we cannot make this large number of measurements, for it would be very impractical for the training of the classifier. In addition, the estimated values of the covariance matrices would be rather inaccurate if the feature dimensionality is too large.

This problem can be solved by the following steps prior to classification. Firstly, we project all the data into a whitened PCA (Principal Component Analysis) space, such that Σ_T becomes an identity matrix with a lower dimensionality of N_{PCA} . At this point, we make a simplifying assumption that each subject shares the same within-class covariance matrix with each other, so that it can be estimated more accurately from the data of all the subjects. It was proved in [2], that in this new feature space, the number of modes of variations contributing to the verification, is not more than $N_{user} - 1$, where N_{user} is the number of subjects for training. Besides, these modes of variations have the smallest variances of each individual subject's data. A further dimensionality reduction can then be achieved by applying another PCA to the data, and discarding all the modes of variations except the $N_{user} - 1$ ones with the smallest variances. This last operation is in fact a dimensionality reduction by means of the LDA (Linear Discriminant Analysis). The whole procedure of dimensionality reduction can be represented by a transformation matrix \mathbf{F} . After the LDA, the total covariance matrix becomes an identity matrix, while the within-class covariance matrix becomes diagonal. Both of them have a dimensionality of $N_{user} - 1$ [3]. As a result, (1) can be rewritten as

$$\begin{aligned} S(\mathbf{x}) = & -(\mathbf{Fx} - \mathbf{F}\mu_W)^T \Lambda_w^{-1} (\mathbf{Fx} - \mathbf{F}\mu_W) \\ & + (\mathbf{Fx} - \mathbf{F}\mu_T)^T (\mathbf{Fx} - \mathbf{F}\mu_T), \end{aligned} \quad (2)$$

where Λ_w denotes the resulting diagonal, within-class covariance matrix. Equation (2) shows that four entities in total need to be estimated from the training data: μ_W , μ_T , \mathbf{F} , and Λ_w .

2.2 Support Vector Machine

The SVM is a binary classifier that maximizes the margin between two classes. Attributed to this characteristic, the generalization performance (i.e. error rates on test sets) of SVM usually either matches or is significantly better than that of competing methods [9], [10].

Given a training set of instance-label pairs $(\mathbf{x}_i, y_i), i = 1, \dots, l$ where $\mathbf{x}_i \in R^n$ and $y_i \in \{1, -1\}^l$, SVM requires the solution of the following optimization problem:

$$\min_{w,b,\xi} \quad \frac{1}{2} w^T w + C \sum_{i=1}^l \xi_i \quad (3)$$

$$\text{subject to } y_i(w^T \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad (4)$$

$$\xi_i \geq 0. \quad (5)$$

Here training vectors \mathbf{x}_i are mapped into a higher (maybe infinite) dimensional space by the function ϕ . Then SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. $C > 0$ is the penalty parameter of the error term, a larger C corresponding to assigning a higher penalty to errors. Furthermore, $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ is called the kernel function.

We applied SVM to the grip-pattern verification, where multiple users are involved, by using the method proposed by [11]. Specifically, the problem is formulated in a difference space, which explicitly captures the dissimilarities between two grip-pattern images. In this difference space, we are interested in the following two classes: the dissimilarities between images of the same individual, and the dissimilarities between images of different people. These two classes are the input to a SVM algorithm, and the SVM algorithm generates a decision surface separating the two classes.

The data are transformed by both PCA and LDA in exactly the same way as in the case of the LRC, prior to the classification. In SVM we used the Gaussian radial basis function kernel.

3 Experiments, Results and Discussion

We recorded the grip-pattern data from a group of police officers in three sessions, with approximately one month and four months in between. In total, 39 subjects participated in both the first and the second collection sessions with 25 grip-pattern images recorded for each subject. In the third session, however, the data were collected from 22 subjects out of the same group, and each subject contributed 50 images. The verification performance is evaluated by the overall EER of all the subjects. It is computed from the matching scores of all the genuine users and impostors.

The experimental results obtained by using the SVM and the LRC are compared in five cases. In the first case, none of the three methods of TMR, DTM and LABP described in Section 1 is in use (see Table 1). Only one of these methods is applied in the second, third, and fourth case, respectively (see Table 2, 3, and 4). In the last case, all of the three methods are implemented (see Table 5).

Table 1. Original verification results in EER(%)

Train	Test	LRC	SVM
1	3	24.09	17.55
2	3	18.95	14.73
1	2	7.94	4.36
3	2	20.16	11.64
2	1	5.53	4.00
3	1	14.70	11.45
Average		15.2	10.6

One can see from Table 1 that if none of the methods of TMR, DTM and LABP is applied, the verification results based on SVM are much better on average, than those

Table 2. Verification results in EER(%) with data preprocessed by TMR

Train	Test	LRC	SVM
1	3	18.36	10.55
2	3	18.88	13.24
1	2	5.98	5.84
3	2	17.82	15.45
2	1	3.90	4.07
3	1	12.91	10.73
Average		12.9	9.9

Table 3. Verification results in EER(%) with DTM

Train	Test	LRC	SVM
1+2	3	13.73	12.73
1+3	2	5.09	5.27
2+3	1	4.00	4.47
Average		7.6	7.4

obtained by LRC. This suggests that SVM is more capable in coping with large across-session variations of data, compared to LRC. This may be attributed to the different characters of these two classifiers. Since the SVM tries to maximize the margin between different classes, it seems to have better generalization performance compared to the LRC, which is based on the PDF estimation of data.

However, Table 2, 3 and 4 show that if one of TMR, DTM or LABP is applied, the SVM is not as much superior to the LRC as in Table 1, even though the verification results based on both classifiers become improved on average. That is, LRC benefits more from these methods than SVM. It is quite interesting to note that LRC actually outperforms SVM, if all three preprocessing methods are combined in use (see Table 5). To summarize, SVM seems to lose its superiority to LRC once the data are better modelled in the training session. What's more, the better the data are modelled, the more superiority SVM tends to lose. Since the best verification results are obtained by using LRC (see Table 5), we shall continue our future work based on LRC, instead of SVM.

Table 4. Verification results in EER(%) with data preprocessed by LABP

Train	Test	LRC	SVM
1	3	9.98	6.91
2	3	16.73	9.81
1	2	5.23	3.09
3	2	11.91	10.31
2	1	4.82	3.57
3	1	8.55	8.76
Average		9.5	7.0

Table 5. Verification results in EER(%) with TMR, DTM and LABP

Train	Test	LRC	SVM
1+2	3	4.86	7.64
1+3	2	3.64	5.82
2+3	1	2.02	3.67
Average		3.5	5.7

4 Conclusions

The grip-pattern verification has been implemented based on both classifiers of LRC and SVM, and the results have been compared under different conditions. It has been shown that SVM gives much better results than LRC, when there are considerable data variations between training and testing. That is, in the situation where data are improperly modelled during the training process, SVM seems to be able to capture the characteristics of data better than LRC. However, once the variations are reduced and thus the data are better modelled during the training process, SVM tends to lose its superiority. Besides, the better the data are modelled, the less SVM tends to outperform LRC.

Acknowledgments

This research is supported by the Technology Foundation STW, applied science division of NWO and the technology programme of the Ministry of Economic Affairs.

References

1. The national Uniform Crime Reporting Program, Law Enforcement Officers Killed and Assaulted, tech. rep., Federal Bureau of Investigation (2001)
2. Shang, X., Veldhuis, R.N.J., Bazen, A.M., Ganzevoort, W.P.T.: Algorithm Design for Grip-Pattern Verification in Smart Gun. In: ProRISC (2005)
3. Veldhuis, R., Bazen, A., Kauffman, J., Hartel, P.: Biometric verification based on grip-pattern recognition. In: SPIE (2004)
4. Shang, X., Veldhuis, R.N.J.: Registration of Hand-Grip Pattern in Smart Gun. In: ProRISC (2006)
5. Jain, A.: Fundamentals of digital image processing. Prentice-Hall, Englewood Cliffs (1989)
6. Shang, X., Veldhuis, R.N.J.: Local Absolute Binary Patterns as Image Preprocessing for Grip-Pattern Recognition in Smart Gun. In: BTAS (2007)
7. Bazen, A., Veldhuis, R.N.J.: Likelihood-Ratio-Based Biometric Verification. IEEE Transactions on circuits and systems for video technology 14(1) (2004)
8. Van Trees, H.L.: Detection, Estimation, and Modulation Theory. Wiley, New York (1968)
9. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, Heidelberg (1995)
10. Burges, C.J.C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2, 121–167 (1998)
11. Jonathon Phillips, P.: Support Vector Machines Applied to Face Recognition. In: Advances in Neural Information Processing Systems, vol. 11. MIT Press, Cambridge (1999)

Chirplet Transform Applied to Simulated and Real Blue Whale (*Balaenoptera musculus*) Calls

Mohammed Bahoura¹ and Yvan Simard²

¹ Département de Mathématiques, d’Informatique et de Génie

Université du Québec à Rimouski,

300, allée des Ursulines, Rimouski, Québec, Canada, G5L 3A1

Mohammed_Bahoura@uqar.qc.ca

² Institut des Sciences de la Mer,

Université du Québec à Rimouski,

310, allée des Ursulines, Rimouski, Québec, Canada, G5L 3A1

Yvan_Simard@uqar.qc.ca

Abstract. Chirplet transform performance to identify low-frequency blue whale calls is tested with simulations and observations from North-West Atlantic. The three different calls are simulated using linear or quadratic frequency sweeping chirps and a hanning window. The performance of Chirplet transform to accurately estimate the chirp parameters with or without noise is first assessed. Then the performance to classify the real vocalizations from the test dataset using the three features best estimated from the simulations is then assessed. The method has a high classification rate and appears promising to efficiently identify these blue whale signature vocalizations with a reduced number of parameters, even under low signal to noise ratios.

Keywords: Blue whale vocalizations, chirplet transform, feature extraction, vector quantization, noise.

1 Introduction

As birds, marine mammals are highly vocalizing animals and different species can be recognized by their specific sounds, calls or songs. Among them, the blue whale produce powerful low-frequency (< 100 Hz) vocalizations that can propagate over large distances, reaching more than 100 km before being masked by the local ocean noise levels, which varies widely in different environments [1][2]. North Atlantic blue whales produce at least three calls that can be used to identify the species in blind recordings at a given location: the stereotyped A and B infrasonic calls (15-20 Hz) and the more variable audible D or arch call (35-120 Hz) [3][4]. Since about 20 years, such signature calls have been used to track the presence or location of various whale species in different ocean basins along their long-range annual migrations routes [5]. Such passive acoustic monitoring (PAM) of whales requires efficient processing algorithms to detect and identify the signature calls on long recording series, extending to several

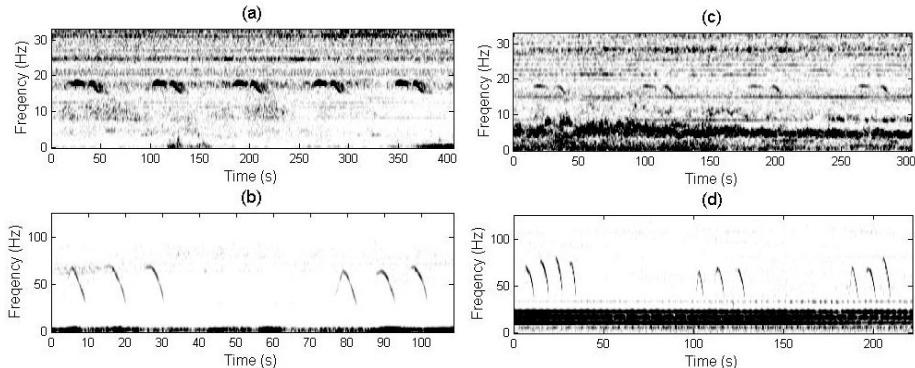


Fig. 1. a) spectrogram of five AB phrases marked with multipath echo, b) spectrogram of six D calls grouped in trio, c) spectrogram of four AB phrases drowned in noise, and d) spectrogram of ten D calls in presence of intense low-frequency noise

months or years Automatic PAM detection algorithms [6] must be efficient, with high detection rate and low false alarms, robust to variable noise conditions and, ideally, fast enough to allow real-time detection and identification under multi-species and multi-sources conditions. Here we present a new method to efficiently identify blue whale signature calls from a limited number of features, which is tested from a representative subset of vocalizations recorded in the St. Lawrence Estuary, a traditional feeding ground where blue whale migrate in summer and where shipping noise from a major continental seaway represent an additional challenge to automatic detection and identification of low-frequency vocalizations.

2 Chirplet Transform

The chirplet transform (CT) first proposed by Mann and Haykin [7] is an extension of the wavelet transform (WT). It is a time-frequency technique, which decomposes the signal using basis functions, called *Chirplets*. In WT, the basis functions are derived from a single mother wavelet by two operations of dilation and translation [8]. Similarly, the basis of chirplets are derived from a single Gaussian function $g(t) = \pi^{-1/4} \exp(-t^2/2)$ using four operations: scaling, chirping, time and frequency shifting. This leads to a family of chirplets with four adjustable parameters.

$$g_{\tau, f, c, d}(t) = \frac{1}{\sqrt{\sqrt{\pi}d}} e^{-\frac{1}{2}(\frac{t-\tau}{d})^2} e^{j[c(t-\tau)+2\pi f](t-\tau)} \quad (1)$$

where τ , f and c are in \mathbb{R} and d is in \mathbb{R}^+ . The parameters τ , f , c and d represent, respectively, the location in time, the location in frequency, the chirp rate and the duration.

The chirplet transform of a continuous-time signal $x(t)$ is defined by the following inner product:

$$a_{\tau,f,c,d} = \int_{-\infty}^{\infty} x(t) g_{\tau,f,c,d}^*(t) dt \quad (2)$$

where $(*)$ denotes the conjugate complex. The coefficient $a_{\tau,f,c,d}$ represents the signal energy content in a time-frequency region specified by the chirplets [8]. To simplify the notation, the parameter set of the n^{th} chirplet is described by an index set $I_n = \{\tau_n, f_n, c_n, d_n\}$. A given signal $x(t)$ can be approximated as a weighted sum of P chirplets.

$$\hat{x}_P(t) = \sum_{n=1}^P a_{I_n} g_{I_n}(t) \quad (3)$$

where $\hat{x}_P(t)$ is defined as the P^{th} order approximation of the signal.

The optimal estimation of a_{I_n} and I_n corresponding to the decomposition of a signal $x(t)$ into the basis function g_{I_n} can be obtained by the maximum likelihood estimation (MLE) algorithm [8]. In this work, the "DiscreteTFDs" toolbox is used to implement the chirplet transform [9].

3 Application to Simulated Calls

In this section, we apply the CT to simulated A, B and D calls to estimate the parameters under clean and noisy conditions.

3.1 AB Phrases

The A call is simulated using a linear frequency sweeping chirp function $p_{f_A, c_A}(t) = \cos(2\pi(f_A + c_A t)t)$ multiplied by a hanning window $h_{d_A}(t) = \frac{1}{2} + \frac{1}{2} \cos(2\pi t/d_A)$, where f_A is the frequency center, c_A is the chirp rate, and d_A is the duration. In the same manner, the B call is simulated using $p_{f_B, c_B}(t) = \cos(2\pi(f_B + c_B t)t)$ and $h_{d_B}(t) = \frac{1}{2} + \frac{1}{2} \cos(2\pi t/d_B)$ functions. An artificial sound, $x_{AB}(t)$, containing A and B calls, respectively centered on t_{c_A} and t_B , is defined by Eq. 4

$$\begin{aligned} x_{AB}(t) &= x_A(t) + x_B(t) \\ &= p_{f_A, c_A} \left(t - \tau_A + \frac{d_A}{2} \right) h_{d_A}(t - \tau_A) \\ &\quad + p_{f_B, c_B} \left(t - \tau_B + \frac{d_B}{2} \right) h_{d_B}(t - \tau_B) \end{aligned} \quad (4)$$

The values of the simulated chirp parameters τ , f , c , and d are reported in Tab. 1. In this example, the A and B calls are simulated using two chirps centered on $\tau_A = 25$ s and $\tau_B = 40$ s, respectively.

A noisy version, $y_{AB}(t)$, is obtained by adding white noise, $n(t)$, to $x_{AB}(t)$, with a signal to noise ratio (SNR) of 0 dB.

$$y_{AB}(t) = x_{AB}(t) + n(t) \quad (5)$$

CT is applied to these signals to approximate each simulated call by one chirplet. The chirplets obtained from the clean signal, $\hat{x}_{AB,1}(t) = \hat{x}_{A,1}(t) + \hat{x}_{B,1}(t)$, are an approximation of A and B calls with a single chirplet, respectively. Similarly, chirplets are estimated for the noisy versions, $\hat{y}_{AB,1}(t) = \hat{y}_{A,1}(t) + \hat{y}_{B,1}(t)$. Approximation by one chirplet corresponds to $P = 1$ in Eq. 3.

Fig. 2 shows the simulated A and B calls, $x_{AB}(t)$, and their approximation by two chirplets, and their corresponding spectrograms for the clean and noisy conditions. The estimated CT parameter values of the obtained chirplets are given in Tab. II.

The chirplet parameters τ , f and c are well estimated, but the parameter d is systematically negatively biased. Values obtained from noisy version are very close to those obtained from the clean signal. CT approximation appears therefore insensitive to noise.

3.2 D Calls

The D call is simulated using a quadratic frequency sweeping chirp $q_{f_D, c_D}(t) = \cos(2\pi(f_D + c_D t^2)t)$ multiplied by a hanning window $h_{d_D}(t) = \frac{1}{2} + \frac{1}{2} \cos(2\pi t/d_D)$, where f_D is the frequency center, c_D is the chirp rate, and d_D is the duration. A signal $x_D(t)$ containing a D call centered on τ_D is generated by Eq. 6.

$$x_D(t) = q_{f_D, c_D} \left(t - \tau_D + \frac{d_D}{2} \right) h_{d_D}(t - \tau_D) \quad (6)$$

The values of the simulated chirp parameters τ_D , f_D , c_D , and d_D are reported in Tab. 2. In this example, a D call is simulated using one quadratic frequency sweeping chirp centered on $\tau_D = 8.5$ s

A noisy version, $y_D(t)$, is obtained by adding white noise $n(t)$ with SNR=0dB.

$$y_D(t) = x_D(t) + n(t) \quad (7)$$

In this section, we used CT to approximate the noisy signal, $y_D(t)$, by one, two and three chirplets. The resulting signals $\hat{y}_{D,1}(t)$, $\hat{y}_{D,2}(t)$ and $\hat{y}_{D,3}(t)$ correspond respectively to $P=1$, 2 and 3 in Eq. 3.

Fig. 3 shows the noisy signal, $y_D(t)$, its approximation by one to three chirplets, $\hat{y}_{D,1}(t)$, $\hat{y}_{D,2}(t)$ and $\hat{y}_{D,3}(t)$, and their corresponding spectrograms. The three-chirplets solution gives the best approximation.

The CT estimation of chirplet parameters (Tab. 2), indicates again that CT approximation is insensitive to noise. For this call, an approximation by a single chirplet gives a relatively accurate estimation of τ and f , but poor estimates of c and d . An approximation by three chirplets is needed to get relatively accurate estimates of all parameters, which can be chosen on this solution as follows: $\tau_D \simeq \tau_{D2} = 8.72$ s, $f_D \simeq f_{D2} = 70.17$ Hz, $c_D \simeq (f_{D3} - f_{D1})/(\tau_3 - \tau_2)^2 = (61.72 - 74.45)/(9.88 - 7.54)^2 = -2.32$ Hz/s and $d_D \simeq d_{D1} + d_{D2} + d_{D3} = 2.28 + 2.23 + 2.01 \simeq 6.52$ s.

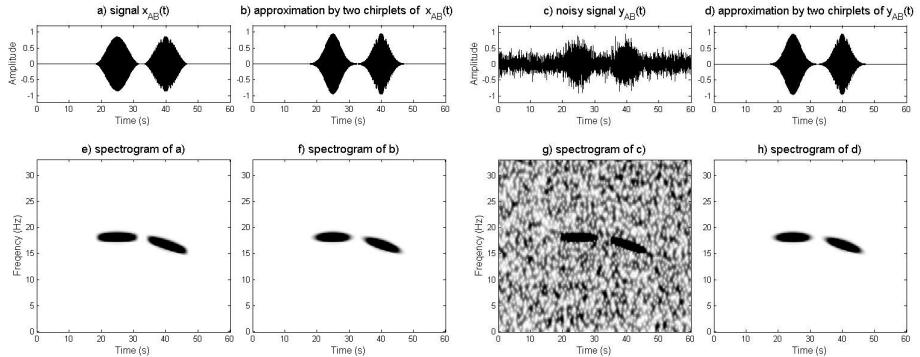


Fig. 2. a) signal $x_{AB}(t)$ representing A and B calls, and b) their approximation by two chirplets, $\hat{x}_{AB,1}(t)$. c) noisy version of A and B calls, $y_{AB}(t)$, with SNR=0dB, and d) its approximation by two chirplets $\hat{y}_{AB,1}(t)$. e-h) represent the corresponding spectrograms.

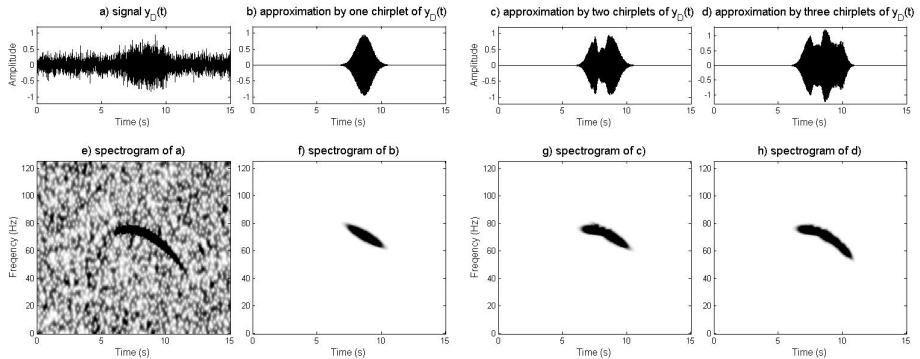


Fig. 3. a) signal representing a noisy simulated D call, $y_D(t)$, using a quadratic frequency sweeping chirp, b) its approximation by one linear frequency sweeping chirplet, $\hat{y}_{D,1}(t)$, c) by two linear frequency sweeping chirplets, $\hat{y}_{D,2}(t)$, and d) by three linear frequency sweeping chirplets, $\hat{y}_{D,3}(t)$, and e-h) represent the corresponding spectrograms

4 Application to Real Calls

In this section, the CT is applied to real blue whale calls to characterize them by a reduced number of parameters.

4.1 Illustrative Example

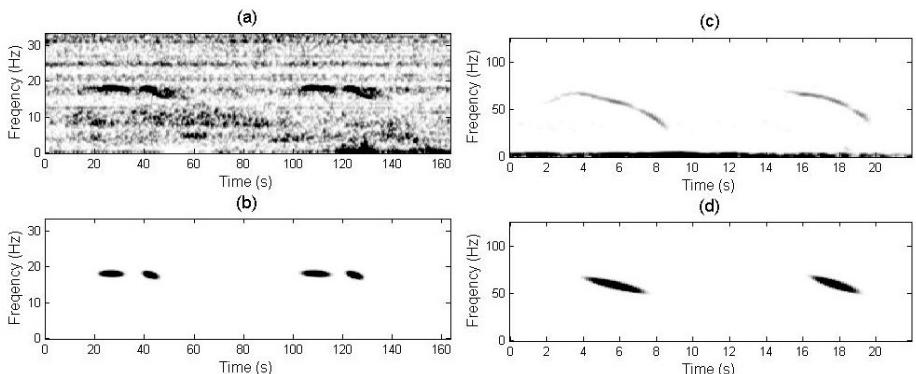
In this example, the AB phrases and D calls are filtered by 13-23 Hz and 40-80 Hz band-pass filters, respectively. Figs. 4-a,b present the spectrogram of a

Table 1. Chirp parameters estimated by a single chirplet for clean and noisy simulated A and B calls

Parameter of chirp	Unit	Original value of simulated signal $x_{AB}(t)$	Estimated value of $\hat{x}_{AB,1}(t)$, on clean signal $x_{AB}(t)$	Estimated value of $\hat{y}_{AB,1}(t)$, on noisy signal $y_{AB}(t)$
A call	τ_A	s	25.00	25.00
	f_A	Hz	18.00	18.00
	c_A	Hz/s	0.00	0.00
	d_A	s	14.00	10.46
B call	τ_B	s	40.00	40.00
	f_B	Hz	16.50	16.50
	c_B	Hz/s	-0.21	-0.21
	d_B	s	14.00	10.46

Table 2. Chirp parameters estimated from multiple chirplets on clean and noisy simulated D call

Parameter of simulation	Unit	Original value of simulated signal $x_D(t)$	Estimated value on clean signal $x_D(t)$	Estimated value on noised signal $y_D(t)$
Approximation by one chirplet: $\hat{x}_{D,1}(t)$ and $\hat{y}_{D,1}(t)$				
τ_D	s	8.50	8.63	8.67
f_D	Hz	72.50	70.49	70.25
c_D	Hz/s	-1.60	-5.33	-5.36
d_D	s	7.00	2.73	2.69
Approximation by two chirplets: $\hat{x}_{D,2}(t)$ and $\hat{y}_{D,2}(t)$				
τ_D	s	8.50	7.543 8.79	7.50 8.4
f_D	Hz	72.50	74.46 69.69	74.54 69.97
c_D	Hz/s	-1.60	-1.08 -6.04	-0.97 -6.03
d_D	s	7.00	2.28 2.62	2.11 2.60
Approximation by three chirplets: $\hat{x}_{D,3}(t)$ and $\hat{y}_{D,3}(t)$				
τ_D	s	8.50	7.54 8.72 9.88	7.51 8.72 9.84
f_D	Hz	72.50	74.45 70.17 61.72	74.52 70.15 62.06
c_D	Hz/s	-1.60	-1.19 -5.63 -9.78	-1.09 -5.63 -9.66
d_D	s	7.00	2.28 2.23 2.01	2.12 2.20 1.61

**Fig. 4.** a) spectrogram of two AB phrases and b) the spectrogram of their approximation with four chirplets. c) Spectrogram of two D calls and d) the spectrogram of their approximation with two chirplets.

signal containing two AB phrases and their corresponding approximation by four chirplets, one chirplet per call. Figs. 4-c,d show the spectrogram of a signal containing two D calls and their corresponding approximation by two chirplets. The calls are accurately detected despite the presence of noise. Approximation by a single chirplet is sufficient for detecting A and B calls, and is acceptable for D calls.

4.2 Database

The test dataset is constructed from sounds recorded in the St. Lawrence Estuary. The blue whale vocalizations were extracted manually and sorted into three classes: A, B and D. Each class contains 100 calls.

4.3 Feature Extraction

Each individual A, B and D vocalization is approximated by one chirplet as in Fig. 4. The feature vector characterizing an individual vocalization is constructed using the location in frequency f , the chirp rate c and the duration d of the corresponding chirp.

$$v = [f, c, d] \quad (8)$$

where f , c and d are expressed in Hz, Hz/s and s, respectively.

4.4 Vector Quantization

In this project, we used the vector quantization (VQ) to classify the blue whale calls. The VQ is a process of mapping vectors from large space to a finite number of regions in that space. Each region is called a cluster and can be represented by its centroid called *codeword*. For a given vocalization class, the resulting codewords constitute the *codebook* of this class. There are two phases in this classification system: training and recognition. In the training phase, an acoustical model (codebook) is constructed for each vocalization class and the models are stored in a database. In the recognition phase, the unknown vocalization is analyzed and the best matching model is searched from the database. In this experience, we used eight codewords per class.

4.5 Results

To overcome the limited number of calls of the test dataset, the “*k-fold cross-validation*” method is employed to evaluate the performance of the VQ-based classifier. The principle consists in dividing each class into 10 groups, each time we take 9 groups for VQ training and the resting one for the classification test. For each vocalization class, the classification performance is evaluated by the ratio of the number of the correctly classified vocalization (NCCV) to the number of the tested vocalizations (NTV) of this class.

$$\text{Performance (\%)} = \frac{\text{NCCV}}{\text{NTV}} \times 100 \quad (9)$$

Table 3. Performance of VQ-based classifier using three chirplet parameters

Vocalization	Performance (%)
A	93.0
B	97.0
D	98.0
Total	96.0

Results are presented in Table 3. The classification rate was high, with 96% of the calls correctly classified.

5 Conclusion

We presented a new approach based on CT to characterize the blue whale calls. The CT decomposes the signal into gaussian chirplet basic functions with four adjustable parameters, i.e., location in time, location in frequency, chirp rate and duration. Results obtained on simulated and real blue whale calls show that this technique is very promising for efficiently detecting and classifying these calls and robust to noise conditions.

References

1. Sirovic, A., Hildebrand, J.A., Wiggins, S.M.: Blue and fin whale call source levels and propagation range in the southern ocean. *J. Acoust. Soc. Am.* 122, 1208–1215 (2007)
2. Stafford, K.M., Mellinger, D.K., Moore, S.E., Fox, C.G.: Seasonal variability and detection range modeling of baleen whale calls in the Gulf of Alaska, 1999–2002. *J. Acoust. Soc. Am.* 122, 3378–3390 (2007)
3. Mellinger, D.K., Clark, C.W.: Blue whale (*Balaenoptera musculus*) sounds from the North Atlantic. *J. Acoust. Soc. Am.* 114(2), 1108–1119 (2003)
4. Berchok, C.L., Bradley, D.L., Gabrielson, T.B.: St. Lawrence blue whale vocalizations revisited: Characterization of calls detected from 1998 to 2001. *J. Acoust. Soc. Am.* 120, 2340–2354 (2006)
5. Mellinger, D.K., Stafford, K.M., Moore, S.E., Dziak, R.P., Matsumoto, H.: An overview of fixed passive acoustic observation methods for cetaceans. *Oceanography* 20(4), 36–45 (2007)
6. Mellinger, D.K., Clark, C.W.: Recognizing transient low-frequency whale sounds by spectrogram correlation. *J. Acoust. Soc. Am.* 107(2), 3518–3529 (2000)
7. Mann, S., Haykin, S.: Adaptive Chirplet Transform: an adaptive generalization of the wavelet transform. *Optical Engineering* 31(6), 1243–1256
8. Cui, J., Wong, W.: The adaptive chirplet transform and visual evoked potentials. *Biomedical Engineering, IEEE Transactions on* 53(7), 1378–1384 (2006)
9. J.C. O'Neill, DiscreteTDFs: Time-Frequency Analysis Software,
<http://tfd.sourceforge.net/>

Illumination Invariant Face Recognition under Various Facial Expressions and Occlusions

Tiwuya H. Faaya and Önsen Toygar

Computer Engineering Department, Eastern Mediterranean University,
Gazimağusa, Northern Cyprus, Mersin 10, Turkey
{faaya.tiwuya, onsen.toygar}@emu.edu.tr

Abstract. This paper presents a new approach that increases face recognition performance using various facial expressions in the presence of illumination variations and occlusions. The new approaches use PCA and LDA with the combination of the preprocessing techniques of histogram equalization and mean-and-variance normalization in order to nullify the effect of illumination changes and any occlusions present which are known to significantly degrade recognition performance. To be consistent with the research of others, our work has been tested on the JAFFE database and its performance has been compared with traditional PCA and LDA.

Keywords: face recognition, facial expressions, face occlusion, illumination, PCA, LDA.

1 Introduction

Face recognition has been extensively studied in recent years. However, it is still an unsolved problem under varying conditions such as different facial expressions, illumination variations and partial occlusions. Researchers studying in this field [1-5] are trying to find robust techniques which recognize faces with different facial expressions, partial occlusions and illumination variations.

A facial expression results from one or more motions of the muscles of the face. These movements can depict sadness, anger, happiness, fear, disgust or surprise on the face of an individual. From the perspective of facial recognition, a facial expression can be considered to consist of deformation of facial components and their spatial relation, or changes in pigmentation of the face [6]. On the other hand, a person may have a medical operation on his/her face and a region on the face can be occluded by a band. These occlusions may randomly appear in any region of the face with different sizes or any specific area of the face such as one of the eyes, both left and right eyes, nose and mouth may also be occluded. In addition to these difficulties, there may be some illumination variations on the facial images in which the faces become darker or brighter according to some uncontrolled environmental conditions.

Each of these difficulties was studied separately to find robust face recognition methods which are not affected by illumination variations, occlusions and facial expressions. In [7], an occlusion robust face recognition scheme with dynamic similarity features was proposed which randomly crops different size patches on the test images.

They used FERET database for the experiments. The patches were small squares which were placed on the cropped face randomly. The proposed method which was a generalization of KDA algorithm was robust to random occlusions to some extent.

In [8], the authors performed experiments on AR database to recognize faces under different conditions such as using partially occluded faces with sunglasses and scarf, and using synthetic occlusion patterns (occlusions on local patches of the facial images). In this context, their S-LNMF based recognition algorithm is better than the other algorithms on real occlusions (sunglass and scarf occlusions) and it performs very well on synthetic occlusions.

Another technique was proposed in [9] which uses Lophoscopic PCA. The authors used six subsets of images in the training set which correspond to the whole face, and the masking of the left eye, right eye, both eyes, nose and mouth respectively. Then they combined these subjects using different combination strategies. They demonstrated that Lophoscopic PCA performs better than PCA approach. They created and used UPC database in their experiments.

On the other hand, [10] presents another scheme for face recognition in the presence of expression and/or illumination variation. Their wavelet-based face recognition scheme demonstrated an improved accuracy compared to PCA. They used JAFFE and ORL databases.

In this study, we considered all the difficulties faced in solving the face recognition problem in the presence of illumination variations, facial expressions and occlusions. These three difficulties were not considered altogether in the same study before and therefore, a new approach is proposed to reduce the effect of these difficulties on the face recognition problem using PCA-based and LDA-based approaches with the combination of the preprocessing techniques histogram equalization (HE) and mean-and-variance normalization (MVN) on the JAFFE database. Various experiments were performed to test the recognition performance of the proposed normalized approaches (nPCA and nLDA) under different illumination variations, six different facial expressions and specific and random occlusions on the facial images.

The rest of the paper is organized as follows. Section 2 presents the details of the proposed approaches. In Section 3, the experimental results for nPCA, nLDA, PCA and LDA approaches with different scenarios on the JAFFE database are demonstrated. Finally, Section 4 presents the conclusions.

2 Our Work

As we have aforementioned, in order to reduce the effect of illumination variations, facial expressions and occlusions altogether for the solution of the face recognition problem, we proposed a new approach using PCA- and LDA-based methods. The new approaches use PCA and LDA with the combination of the preprocessing techniques histogram equalization (HE) and mean-and-variance normalization (MVN). Both HE and MVN normalization are used to improve the face recognition performance since feature normalization techniques are able to largely reduce the actual mismatch between training and testing conditions [11]. In this study, the JAFFE database is used with various occlusion types applied to specific and random regions on the human face. An example subject from JAFFE database [12] with neutral face image

and the six facial expressions (happy, disgust, fear, surprise, sad, angry) are demonstrated in Fig. 1. Specific occlusions include the regions on the face such as the left eye, right eye, both eyes, nose and mouth as shown in Fig. 2. Random occlusions are arbitrarily chosen patches with different size (dimension) and position on the facial image. Fig. 3 is a demonstration of random occlusions with five different size patches randomly placed on the cropped facial images.



Fig. 1. An example subject from JAFFE database with neutral expression and six facial expressions (angry, happy, sad, surprise, disgust, fear)

The aim of this study is to eliminate the effect of illumination variations and occlusions on facial images that has different facial expressions. The difficulty of face recognition under different facial expressions is known. In addition to this, occlusions are added on the corresponding facial images which include the following facial expressions: happiness, disgust, fear, surprise, sadness, anger.

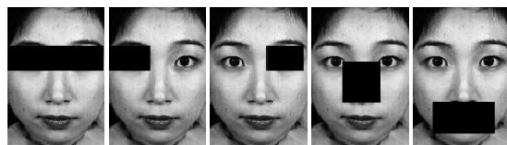


Fig. 2. Specific occlusions on both eyes, left eye, right eye, nose and mouth region on the face

In this study, we first reduce the effect of illumination variation using histogram equalization (HE) and mean-and-variance normalization (MVN) [11] as a preprocessing technique. Then, in order to reduce the effect of facial expressions and occlusions on the face recognition performance, we applied PCA- and LDA-based normalized approaches (nPCA and nLDA) on the cropped facial images. In all recognition experiments, Euclidean distance measure was used for classification. Different scenarios on how to apply occlusions on the facial images were studied. The first scenario puts the occlusions on the neutral training images and the remaining six facial expressions for each person were tested using both normalized and traditional approaches. Then, occlusions were applied on the test images which were demonstrating one of the six facial expressions. The final scenario, which is a new scheme, is to apply occlusions on both test and training images in order to improve the recognition performance. The experimental results on these scenarios are demonstrated in the next section with different occlusion types.

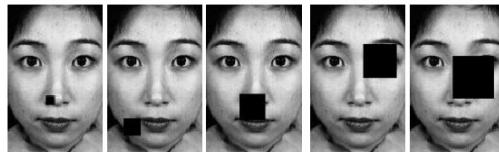


Fig. 3. Random occlusions with five different size patches randomly placed on the cropped facial images (10x10, 20x20, 30x30, 40x40, 50x50 patches)

3 Experimental Results

Recognition performance of the new approach is measured using the JAFFE database. We performed 3 sets of experiments on the JAFFE database to compare the face recognition performance of the new PCA-based and LDA-based approaches (nPCA and nLDA) and the traditional PCA and LDA.

Japanese Female Facial expression Database (JAFFE) contains 213 images of 10 Japanese female facial expressions with 3 or 4 examples of each of the 6 basic facial expressions (happy, sad, surprise, anger, disgust, fear) and a neutral face. Original 256x256 images were cropped automatically by aligning the eyes at the same position and the images were rescaled down to 116x180. In each experimental set, 2 neutral images of each subject from the database is chosen and used as training images. Two sample images from the remaining 6 facial expressions of each subject are chosen to test the recognition performance of the methods used in the experiments.

In the first set of experiments the effect of illumination is studied on the face recognition problem under different facial expressions. We brighten and darken the test images by 5, 10, 15, 20, 25 percent and evaluate the recognition performance of PCA, LDA, nPCA and nLDA methods. The results of the brightening and darkening operations are demonstrated in Table 1 and Table 2. It is clear that nPCA performs better compared to the original PCA but LDA performs better than nLDA whenever the images become brighter. Darkening the images reduces the performance of PCA significantly whereas nPCA performs well in this situation. On the other hand, LDA and nLDA records a slight drop in performance whenever we darken the images.

In general, nPCA is not sensitive up to 25% illumination (either brightening or darkening) whereas PCA is sensitive to these variations. For brightening, LDA is not sensitive up to 20% whereas nLDA is sensitive to this type of changes. However they (LDA and nLDA) are both sensitive to darkening since we see a slight drop in their performance.

The second set of experiments evaluate the recognition performance of the methods under specific occlusions on both eyes, left eye, right eye, nose and mouth regions of the facial images. The experiments in this set are performed in 3 scenarios such that the specific occlusions are placed on the training images, on the test images, and on both of them. The results are presented in Table 3 through Table 5 for PCA, nPCA, LDA and nLDA.

The recognition performance of all the methods in Table 3 indicates that specific occlusions applied on the training images only do not affect the recognition performance in the presence of facial expressions on the test images.

Table 1. Recognition Rates (%) under brightened test images

Brightening Percentage	PCA	nPCA	LDA	nLDA
0	80.83	85	88.33	88.33
5	76.66	85	88.33	84.16
10	59.16	85	88.33	84.16
15	34.16	85	88.33	84.16
20	25	85	88.33	84.16
25	20.83	85	87.5	85

Table 2. Recognition Rates (%) under darkened test images

Darkening Percentage	PCA	nPCA	LDA	nLDA
0	80.83	85	88.33	88.33
5	67.5	84.16	87.5	85
10	40.83	84.16	85.83	82.5
15	24.16	84.16	83.33	81.66
20	18.33	85	80	80
25	11.66	85	74.16	80

The recognition performance over the facial expressions and occlusions on the test images of LDA and nLDA are similar and these methods are not very sensitive to specific occlusions. However, PCA and nPCA are both sensitive to specific occlusions and facial expressions over the test images with nPCA performing better than PCA as shown in Table 4.

On the other hand, whenever we apply occlusion on both training and test images, we realize from Table 5 that LDA and nLDA perform similarly and they are not very sensitive to occlusions in this scenario. PCA and nPCA are also not sensitive but nPCA is slightly better than PCA over the specific occlusion on both the training and test images.

Table 3. Recognition Rates (%) with Specific Occlusions on Training Images

Occlusion Region	PCA	nPCA	LDA	nLDA
None	80.83	85	88.33	88.33
Both eyes	80.83	83.33	84.17	85
Left eye	80	82.5	79.17	87.5
Right eye	80.83	84.17	66.67	80.83
Nose	77.5	84.17	57.5	84.17
Mouth	79.17	84.17	72.5	82.5

Table 4. Recognition Rates (%) with Specific Occlusions on Test Images

Occlusion Region	PCA	nPCA	LDA	nLDA
None	80.83	85	88.33	88.33
Both eyes	30	58.33	83.33	82.5
Left eye	59.17	75	80	82.5
Right eye	72.5	73.33	88.33	82.5
Nose	40.83	55.83	71.67	73.33
Mouth	37.5	71.67	65.83	80.33

Table 5. Recognition Rates (%) with Specific Occlusions on both Training and Test Images

Occlusion Region	PCA	nPCA	LDA	nLDA
None	80.83	85	88.33	88.33
Both eyes	80.83	83.33	84.16	85.83
Left eye	80	83.33	85.83	88.33
Right eye	80.83	84.16	78.33	86.66
Nose	77.5	82.5	80.83	87.5
Mouth	79.16	84.16	88.33	87.5

In general if we have specific occlusions on the training images, all the methods behave similarly and they are not affected much by these occlusions. However, putting specific occlusions on the test images does not affect the performance of LDA and nLDA while a performance decrease occurs for PCA and nPCA with PCA showing the greater degree of sensitivity. For the case of occlusions on both training and test images, the recognition performance for PCA and nPCA are not affected that much, nevertheless, nPCA performs better than PCA.

The effect of random occlusion is evaluated using 3 scenarios by applying patches of various sizes (dimensions) at arbitrary positions on the facial images. The dimensions of the patches are 10x10, 20x20, 30x30, 40x40 and 50x50. The third set of experiments is performed by repeating each experiment 3 times and then taking the

Table 6. Recognition Rates (%) with Random Occlusions on Training Images

Patch Dimension	PCA	nPCA	LDA	nLDA
None	80.83	85	88.33	88.33
10x10	80.83	84.72	83.61	87.77
20x20	80.83	79.17	82.5	86.67
30x30	64.97	62.77	65.55	65
40x40	41.66	33.05	46.38	53.05
50x50	41.66	20.81	46.94	47.22

Table 7. Recognition Rates (%) with Random Occlusions on Test Images

Patch Dimension	PCA	nPCA	LDA	nLDA
None	80.83	85	88.33	88.33
10x10	76.94	84.72	88.33	87.22
20x20	74.17	83.33	86.67	84.17
30x30	70.27	79.72	84.72	82.22
40x40	58.33	76.94	83.05	81.11
50x50	43.32	73.05	85.27	72.22

Table 8. Recognition Rates (%) with Random Occlusions on both Training and Test Images

Patch Dimension	PCA	nPCA	LDA	nLDA
None	80.83	85	88.33	88.33
10x10	78.87	85	86.22	84.17
20x20	74.58	67.78	73.13	78.12
30x30	58.05	55.64	51.04	62.92
40x40	35.55	29.44	26.25	32.5
50x50	23.75	17.44	25	28.75

average of these 3 runs. The experimental results demonstrated in Table 6 through Table 8 are the average recognition rates of the 3 runs.

As shown in the above tables, whenever random occlusions are placed on the training images, we see that all methods are sensitive to this type of occlusion with nLDA only slightly outperforming LDA as the patch size increases. Meanwhile PCA performs better than nPCA as the patch size increases.

On the other hand whenever we have random occlusions on the test images, the situation is reversed. LDA performs better than nLDA as the patch size increases while nPCA performs better than PCA. In general, all methods are sensitive to random occlusions applied on the test images. For the final scenario in which occlusions are applied on both training and test images, all the methods show a greater degree of sensitivity than other scenarios, there is a sharp decrease in performance as the patch sizes increases. In that case, PCA and nLDA show slightly better performances than nPCA and LDA respectively.

4 Conclusions

This paper has presented a new approach that increases face recognition performance over the traditional approaches (PCA and LDA) using facial expressions in the presence of illumination variations and occlusions. The results obtained has shown that nPCA gives a better performance over PCA in all the scenarios considered with the exception of when there is a random occlusion on both training and test images. For LDA-based approaches when there is a specific occlusion on a training set or when

specific occlusions are present on both training and test set, nLDA should be preferred to LDA. Also when random occlusions are present on training images or on both training and test images, nLDA outperforms LDA. Either LDA or nLDA could be used in situations when the test images are darker than training images or with specific occlusions on test images. However LDA outperforms our approach as illumination increases or with the presence of random occlusions only on the test images.

References

1. Martinez, A.M.: Recognizing Imprecisely Localized, Partially Occluded and Expression Variant Faces from a Single Sample per Class. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 24, 748–763 (2002)
2. Kim, J., Choi, J., Yi, J., Turk, M.: Effective Representation Using ICA for Face Recognition Robust to Local Distortion and Partial Occlusion. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 27, 1977–1981 (2005)
3. Toygar, Ö.: Sensitivity Analysis of Partitioning-based Face Recognition Algorithms on Occlusions. In: Proceedings of the 6th WSEAS International Conference on Applications of Electrical Engineering (AEE 2007), İstanbul, Turkey, May 2007, pp. 21–27 (2007)
4. Sahbi, H., Boujemaa, N.: Robust face recognition using Dynamic Space Warping. In: Proceedings of ECCV's Workshop on Biometric Authentication, pp. 121–132 (2002)
5. Martinez, A.M.: Recognizing expression variant faces from a single sample image per class. In: Proceedings of 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 2003, vol. 1, pp. 353–358 (2003)
6. Fabrice, B., Claude, C.C., Adrian, A.L.: Recognition of Facial Expressions in the Presence of Occlusion. *BMVC* (2001)
7. Liu, Q., Yan, W., Lu, H., Ma, S.: Occlusion Robust Face Recognition with Dynamic Similarity Features. In: 18th International Conference on Pattern Recognition (ICPR 2006), Hong Kong, pp. 544–547 (2006)
8. Oh, H.J., Lee, K.M., Lee, S.U.: Occlusion invariant face recognition using selective LNMF basis images. In: ACCV (Asian Conference on Computer Vision) 2006, Hyderabad, India, January 2006, pp. 120–129 (2006)
9. Tarrés, F., Rama, A., Torres, L.: A Novel Method for Face Recognition under Partial Occlusion or Facial Expression Variations. In: 47th International Symposium ELMAR-2005, Multimedia Systems and Applications, Zadar, Croatia (June 2005)
10. Harin, S., Sabah, J.: Face Recognition in the Presence of Expression and/or Illumination Variation Fourth. In: IEEE Workshop on Automatic Identification Advanced Technologies (AutoID 2005), pp. 144–148 (2005)
11. Pujol, P., Macho, D., Nadeu, C.: On Real-Time Mean-and-Variance Normalization of Speech Recognition Features. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2006 (May 2006)
12. Michael, J.L., Shigeru, A., Miyuki, K., Jiro, G.: Coding Facial Expressions with Gabor Wavelets. In: Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition, April 1998, pp. 200–205. IEEE Computer Society, Nara, Japan (1998)

A Version of the FastICA Algorithm Based on the Secant Method Combined with Simple Iterations Method

Doru Constantin and Luminita State

University of Pitesti
Department of Computer Science
Address Str. Tg. din Vale, No.1, Romania
cdomanid@yahoo.com, radus@sunu.rnc.ro

Abstract. The work proposes a new algorithm for the estimation of the ICA model, an algorithm based on secant method and successive approximations. The first sections briefly present the standard FastICA algorithm based on the Newton method and a new version of the FastICA algorithm. The proposed algorithm to estimate the independent components combines the secant iterations with successive approximations technique. The final section presents the results of a comparative analysis experimentally derived conclusions concerning the performance of the proposed method. The tests were performed of several samples of signal files.

Keywords: Independent Component Analysis, Blind Source Separation, Numerical Method.

1 Introduction

An important problem arising in signal processing, mathematical statistical and neural networks is represented by the need of getting adequate representations of multidimensional data. The problem can be stated in terms of finding a function f such that the m dimensional transform defined by $s = f(x)$ possesses some desired properties, where x is a m -dimensional random vector. Being given its computational simplicity, frequently the linear approach is attractive, that is the transform is

$$s = Wx \quad (1)$$

where W is a matrix to be optimally determined from the point of view of a pre-established criterion.

There are a long series of methods and principles already proposed in the literature for solving the problem of fitting an adequate linear transform for multidimensional data [1,4], as for instance, Principal Component Analysis (PCA), factor analysis, projection methods and Independent Component Analysis (ICA).

The aim of Independent Component Analysis is to determine a transform such that the components $s_i, i = \overline{1..n}$ becomes statistically independent, or at

least almost statistically independent. In order to find a suitable linear transform to assure that (II) $s_i, i = \overline{1..n}$ become 'nearly' statistically independent several methods have been developed so far. Some of them, as for instance Principal Component Analysis and factor analysis are second order approaches, that is they use exclusively the information contained by the covariance matrix of the random vector x , some of them, as for instance the projection methods and blind deconvolution are higher order methods that use an additional information to reduce the redundancies in the data. Independent Component Analysis has became one of the most promising approaches in this respect and, consists in the estimation of the generative model of the form $x = As$, where the $s = (s_1 s_2, \dots s_n)^T$ are supposed to be independent, and A is the mixing matrix $m \times n$ -dimensional of the model. The data model estimation in the framework of independent component analysis is stated in terms of a variational problem formulated on a given objective function.

The aim of the research reported in this paper is to introduce a new version of the FastICA algorithm; an algorithm that is based on secant iterations combined with successive approximations and to analyze the performances of the algorithm in signal applications.

2 Fixed-Point ICA Based on Iterations Method and Secant Method

2.1 The Standard FastICA Algorithm

In this section the ICA model and the standard FastICA algorithm are briefly exposed. The ICA model is state as $x = As$, where x is the observations vector and A is the mixing matrix of the original sources, s . The aim is to determine the sources, on the basis of x . One of the basic working assumption in estimation the ICA model is that the sources s are statistically independent and they have nongaussian distributions. This way the problem becomes to find the weighting matrix W (the demixing matrix), such that the transform $y = Wx$ gives suitable approximations of the independent sources.

In the following, the numerical estimation of the independent components is going to be obtained using the secant method combined with the successive approximation approaches, the variational problem being imposed on the negentropy taken as criterion function.

The negentropy is defined by:

$$I(y) = H(y_{gauss}) - H(y) \quad (2)$$

where $H(y) = - \int p_y(\eta) \log p_y(\eta) d\eta$ is the differential entropy of the random vector y .

Being given that the Gaussian repartition is of largest differential entropy in the class of the repartitions having the same covariance matrix, the maximization

of the negentropy (2) gives the best estimation of the ICA model. Although this approaches is well founded from information point of view the direct use of the expression (2) is not computationally tractable, and some approximations are needed instead. We use the approximation introduced in (Hyvärinen, 98):

$$I(y) = [E\{G(y)\} - E\{G(\nu)\}]^2 \quad (3)$$

where G is an nonquadratic function, ν and y are Gaussian variables of zero mean and unit variance. Some of the most frequently used expressions of G are,

$$G_1(y) = \frac{1}{a_1} \log \cosh(a_1 y); \quad 1 \leq a_1 \leq 2, \quad G_2(y) = -\exp(-\frac{y^2}{2}); \quad G_3(y) = \frac{y^4}{4}$$

Note that the expressions of their first order derivatives are given by: $g_1(y) = \tanh(a_1 y)$; $g_2(y) = y \exp(-\frac{y^2}{2})$; $g_3(y) = y^3$, respectively.

The variational problem can be formulated as a constraint optimization problem as follows,

$$\max F(w), \quad \|w\|^2 = 1 \quad (4)$$

that is the objective function $F(w)$ has to be maximized on the unit sphere. In case the negentropy is taken as the objective function, we get,

$$F(w) = [E\{G(y)\} - E\{G(\nu)\}]^2 \quad (5)$$

where $y = w^T z$.

To solve the optimization problem from the (4) relation we write the Lagrange function using the Lagrange multiplications method:

$$L(w) = F(w) - \lambda(\|w\|^2 - 1) \quad (6)$$

The necessary conditions for the critical points are:

$$\frac{\partial L(w, \lambda)}{\partial w} = 0 \quad \text{and} \quad \frac{\partial L(w, \lambda)}{\partial \lambda} = 0 \quad (7)$$

Applying (7) in the (6) relation we have:

$$\frac{\partial L(w, \lambda)}{\partial w} = \frac{\partial F(w)}{\partial w} - 2\lambda w = 0 \quad (8)$$

The expression of the gradient $\frac{\partial F(w)}{\partial w}$ is calculated like this:

$$\begin{aligned} \frac{\partial F(w)}{\partial w} &= \frac{\partial [E\{G(y)\} - E\{G(\nu)\}]^2}{\partial w} = \\ &= 2 * [E\{G(w^T z)\} - E\{G(\nu)\}] * [E\{zg(w^T z)\}] \end{aligned} \quad (9)$$

and $\gamma = [E\{G(w^T z)\} - E\{G(\nu)\}]$ is a constant, because ν is a Gaussian random variable. Thus we obtain:

$$\frac{\partial F(w)}{\partial w} = \gamma [E\{zg(w^T z)\}] \quad (10)$$

and it presents the gradient of the negentropy.

Replacing the (10) relation in (8) we obtain:

$$F^*(w) = E\{zg(w^T z)\} - \beta w = 0 \quad (11)$$

where β is a real constant, $\beta = E\{w_0^T zg(w_0^T z)\}$, where w_0 is the critical value of w .

The Newton method applied to (11) gives

$$w \leftarrow E\{zg(w^T z)\} - E\{g'(w^T z)\}w \quad (12)$$

The weighting vectors being normalized, we arrive at the following approximation scheme,

1. $w \leftarrow E\{zg(w^T z)\} - E\{g'(w^T z)\}w$
2. $w \leftarrow w / \|w\|$.

2.2 The FastICA Algorithm Based on the Simple Iterations Method and Secant Method

In this sections, we proposed a version of the FastICA algorithm based on the secant method combined with successive approximations. Experimentally supported conclusions concerning the performance of these methods are reported in the next section.

A secant method [3] in solving the equation (11) yields to the following iterative scheme:

1. select the initial approximation w_0 and a randomly generated value a .
2. apply the updating rule: $\Delta w = -\frac{F^*(w)}{F^*(w) - F(a)}(w - a)$.
3. if the convergence criterion does not hold then go to the step 2, else take the last value of the w^k as the approximative solution of (11).

The convergence criterion is $\|w^{k+1} - w^k\| < \varepsilon$, where $\varepsilon = 10^{-N}$ is a small real constant, $N \in N^*$ given.

Using the successive approximation method, the approximations sequence can be written as $w = \varphi(w)$ where $\varphi(w) = w - 1/M * F^*(w)$ and M is the maximum value of $F^{*'}(w)$.

The updating rule becomes: $w \leftarrow w - 1/M * [E\{zg(w^T z)\} - \beta w]$, where g is previously defined.

The deducing of the ICA algorithm based on secant iterations combined with successive approximations (SSAM Algorithm). In this section we are developing the SSAM algorithm (Secant Method and Successive Approximations Method- based algorithm) that combines two numerical methods.

Aiming to obtain a better data representation and a better convergence order, we introduced the independent components analysis algorithm (SSAM algorithm) operating by: two iterations are performed by simple iterations method and the next iteration is calculated by secant method. The secant method is applying to the function F defined by $F(w) = w - F^*(w)$ using that two points previously obtained by iterations method. By example, let w_0 the initial iteration. Taken $f(w) = F^*(w)$, $w_1 = f(w_0)$ and $w_2 = f(w_1)$ then w_3 is given by:

$$w_3 = \frac{w_0 F(w_1) - w_1 F(w_0)}{F(w_1) - F(w_0)} \quad (13)$$

Substituting the values of the function F we obtain:

$$w_3 = \frac{w_0 w_2 - w_1^2}{w_0 - 2w_1 + w_2} \quad (14)$$

Starting with w_3 , another two iterations are generated with successive approximations method and the next value is obtained with secant method and then we repeat this procedure. We note that in the sequence $w_0, w_1, w_2, w_3, w_4, w_5, w_6, w_7 \dots$ every element w_{3i} , $i = 0, 1, \dots$, may be written in terms of the w_{3i-3} , where $w_k = f(f(w_{k-2}))$. Denote this sequence by w_0, w_1, w_2, \dots , we obtain the following iterative scheme:

$$w_{k+3} = \frac{w_k f(f(w_k)) - f(w_k)^2}{w_k - 2f(w_k) + f(f(w_k))} \quad (15)$$

As compared to the Newton method, the convergence rate of the iterative method (15) is at least of order two. Thus we get an iterative scheme yielding to an improved convergence rate in estimating the independent components.

SSAM Algorithm - version of the standard algorithm for estimating the independent components

Step 1 : Center the data to mean.

Step 2 : Apply the whitening transform to data.

Step 3 : Select the number of independent components n and set counter
 $r \leftarrow 1$.

Step 4 : Select the initial guess of unit norm for w_r .

Step 5 : Apply the updating rules:

$$w_r^k = \frac{w_r^{k-3} f(f(w_r^{k-3})) - f(w_r^{k-3})^2}{w_r^{k-3} - 2f(w_r^{k-3}) + f(f(w_r^{k-3}))} \quad (16)$$

where w_r^0 is a initial value random generated and $f(w_r^k) = F^*(w_r^k)$, $w_r^{k-1} = f(w_r^{k-2})$, $w_r^{k-2} = w_r^{k-3}$.

- Step 6 : Apply the orthogonalization transform: $w_r \leftarrow w_r - \sum_{j=1}^{r-1} (w_r^T w_j) w_j$
- Step 7 : $w_r \leftarrow w_r / \|w_r\|$.
- Step 8 : If w_r has not converged ($\|w_r^{k+1} - w_r^k\| > \varepsilon$, where ε is a small real constant), go back to step 5.
- Step 9 : Set $r \leftarrow r + 1$. If $r \leq n$ then go to step 4.

3 Experimental Analysis

The assessment of the performances of the proposed algorithm for determining of the independent components is achieved in problems of signals recognition. We define an absolute mean sum error (AbsMSE) for comparing the accuracy of matching between original signals and restored signals. Then AbsMSE is defined as follows:

$$AbsMSE = \sum_{i=1}^N |s_i - s_{estimated_i}| / N \quad (17)$$

where s_i and $s_{estimated_i}$ represent the i -th pixel values for original and restored signals, respectively, and N is the total number of pixels.

All the presented tests comprise the recognition performances of the independent components using as an objective function the negentropy for which they used one at a time in the approximation the three functions adopted in the field [1].

In a comparative study the proposed method based on successive approximations combined with secant iterations has recognition performances of the original signals which are similar with the implemented methods, such as FastICA based on the secant method [2] or standard method.

3.1 Experimentally Derived Conclusions on the Performance of the Algorithm in the Case of the Mixtures of the Signals

Test I. We consider as observation data two signals which are mixed and recorded based on two independent components. In this first test, the original

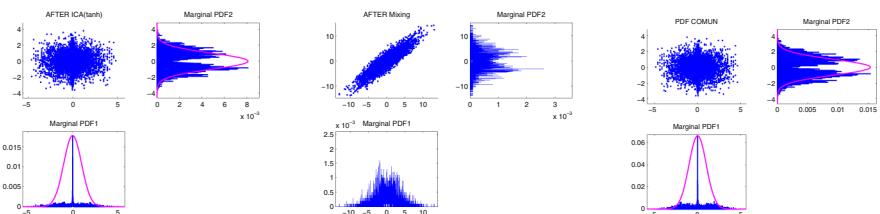


Fig. 1. Source Signals Discovered by the Algorithm (Left: 3 images), The Mixed Signals (Middle: 3 images) and Original Signals (Right: 3 images)

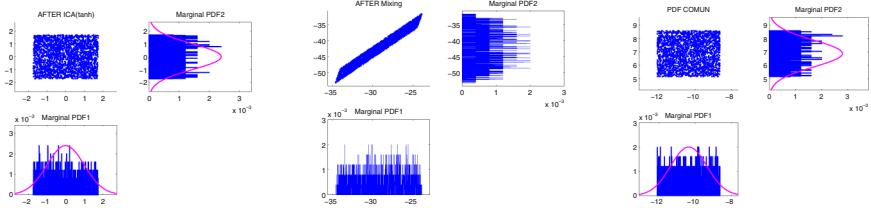


Fig. 2. Source Signals (uniform) Discovered by the Algorithm (Left: 3 images), The Mixed Signals (uniform) (Middle: 3 images) and Original Signals (uniform) (Right: 3 images)

sources are signals generated using the Matlab functions, and the results obtained after applying the SSAM algorithm based on successive approximations and secant iterations show a recognition performance similar to the standard FastICA method based on the Newton and to the method FastICA method based on the secant method [2].

The source signals discovered by the algorithm, the mixed signals and the source signals generated by Matlab subjected to the analysis procedure in independent components are represented in figure 1. In the respective figure we can notice the marginal densities corresponding to the two signals as well as the joint density which is common to the mixtures for the source signals discovered by the algorithm, for the mixed signals and for the source signals, respectively.

The results of the test regarding the appliance of the SSAM algorithm are given in table 1.

Test II. This test resembles the previously test with the difference that it uses, as original signals, the uniform distribution signals.

The figure 2 comprise the original source signals, the mixed signals and the source signals discovered by the algorithm for the uniform signals case.

The results obtained after the comparative study regarding the proposed method and other methods used in the estimation of the ICA model, are similar to the ones from the first test conform with table 1.

Table 1. AbsMSE of versions of the FastICA Algorithm for experimental tests

FastICA	Test I	Test I	Test I	Test II	Test II	Test II	Test III	Test III	Test III
Basic Method	tanh (g_1)	exp (g_2)	kurt (g_3)	tanh (g_1)	exp (g_2)	kurt (g_3)	tanh (g_1)	exp (g_2)	kurt (g_3)
Newton	1.0702* 10^{-2}	1.0706* 10^{-2}	1.0697* 10^{-2}	2.3617* 10^{-2}	2.3166* 10^{-2}	2.4172* 10^{-2}	4.493 * 10^{-3}	4.491 * 10^{-3}	5.813 * 10^{-3}
Secant	1.0716* 10^{-2}	1.0682* 10^{-2}	1.2793* 10^{-2}	2.3271* 10^{-2}	2.2754* 10^{-2}	2.3181* 10^{-2}	5.792 * 10^{-3}	5.741 * 10^{-3}	5.549 * 10^{-3}
SSAM	1.0706* 10^{-2}	1.0695* 10^{-2}	1.0699* 10^{-2}	2.3295* 10^{-2}	2.2981* 10^{-2}	2.2996* 10^{-2}	5.284 * 10^{-3}	5.352 * 10^{-3}	5.382 * 10^{-3}

3.2 Experimentally Derived Conclusions on the Performance of the Algorithm in the Case of Image Signal

Test III. The achieved test refers to the capacity of the proposed algorithm of recognizing independent image faces from images of the mixed faces which can represent joint and superimposed faces as well as deteriorated images subjected to restoration.

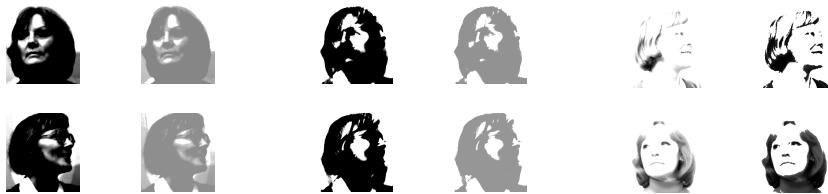


Fig. 3. Original Faces Discovered by the Algorithm (Left: 2 images), The Mixed Faces for Face Recognition (Middle: 2 images) and Original Faces for Face Recognition (Right: 2 images)

In this test we considered again the bidimensional case with two mixed images over which we apply the deterioration algorithm of the original faces. The original image faces, the mixed image faces and the image faces discovered by the proposed algorithm are in figure 3. Just as the previously examples, the obtained results offer good recognition performances of the original faces, showing also in this case a qualitative superiority with respect to convergence speed in the recognition (the results are presented in table 1) compared with other used methods and anterior specified.

4 Summary and Conclusions

In this article we developed an algorithm for estimating the independent components based on a iterative scheme that combines the successive approximations method to the secant method. We derived a suitable algorithm and supplied a comparative analysis of its recognition capacities against the previously developed algorithm. In order to derive conclusive remarks the tests were performed on different signal samples.

References

1. Hyvarinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons, Chichester (2001)
2. Cho, Y.H., Hong, S.J., Park, S.M.: Face Recognition by Using Combined Method of Centroid Shift and Fixed-Point ICA Based on Secant Method. In: Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA 2006), IEEE Computer Society, Los Alamitos (2006)

3. State, L., Cocianu, C., Panayiotis, V., Constantin, D.: PRINCIPAL DIRECTIONS - BASED ALGORITHM FOR CLASSIFICATION TASKS. In: Proceedings of the Ninth International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2007), Timisoara, September 26-29, 2007, IEEE Computer Society, Los Alamitos (2007)
4. Stone, J.V.: Independent Component Analysis. The MIT Press, Cambridge (2004)
5. Bartlett, M.S., Movellan, J.R., Sejnowski, T.J.: Face Recognition by Independent Component Analysis. *IEEE Transactions on Neural Networks* 13 (2002)

SVM-Based Face Recognition Using Genetic Search for Frequency-Feature Subset Selection

Aouatif Amine, Mohammed Rziza, and Driss Aboutajdine

GSCM-LRIT, Faculty of Sciences, Mohammed V University,

B.P. 1014 Rabat, Morocco

aouatif.amine@ieee.org, {rziza, aboutaj}@fsr.ac.ma

Abstract. The goal of this paper is to study if there is a dependency between a selected feature vector at each generation of the genetic algorithm and the resulting fitness. In order to see the relation between these parameters, we first use Discrete Cosine Transforms (DCT) to transform each image as a feature vector (i.e., Frequency Feature Subset (FFS)). A Genetic Algorithm (GA) is then used to select a subset of features from the low-dimensional representation by removing certain DCT coefficients that do not seem to encode important information about recognition task. When using SVM, two problems are confronted: how to choose the optimal input feature subset for SVM, and how to set the best kernel parameters. Therefore, obtaining the optimal feature subset and SVM parameters must occur simultaneously. We present a genetic algorithm approach for feature selection and parameters optimization to solve this kind of problem.

mots clef- Face recognition, Feature Selection, Genetic Algorithm, Support Vector Machine, Discrete Cosine Transform.

1 Introduction

Machine recognition of faces is becoming more and more popular and the need for accurate and robust performance is increasing. Face recognition, as an unsolved problem under the conditions of pose and illumination variations, still attracts significant research efforts. The main reasons for the ongoing research are: (i) the increased need for natural identification for authentication in the networked society, for surveillance, for perceptual user interfaces, (ii) and the lack of robust features and classification schemes for the face recognition task.

A common objective in face recognition is to find a good way of representing face information. High information redundancy present in face images results in inefficiencies when these images are used directly for recognition, identification and classification. A key point in developing a good representation is to expose the constraints and remove the redundancies contained in pixel images of faces. Typically one builds a computational model to transform pixel images into face features, which generally should be robust to variations of illumination, scale and orientation and then use these features for recognition [1]. For classical pattern recognition techniques, the patterns are generally represented as a vector of feature values. The selection of features can have a considerable impact on the effectiveness of the resulting classification algorithm [5]. It is

not often known in advance which features will provide the best discrimination between classes, and it is usually not feasible to measure and represent all possible features of the objects effects. With feature selection, the cost of classification can be reduced by limiting the number of features which must be measured and stored. Some, but not all, feature selection methods realize this benefit as well. A number of approaches for feature subset selection have been proposed in the literature. Koller et al [2] used a greedy algorithm to remove the features that provide the least additional information given the remaining features. Brill et. al [3] have explored randomized population-based heuristic search approaches such as GAs to select feature subsets for NNs. As is known, in many supervised learning problems, feature selection is important, and for SVM, it also performs badly when there are many irrelevant features [4]. In order to improve its performance, suitable feature selection algorithm, such as MLR (Multiple Linear Regression), GA, should be adopted. GAs are good candidates for attacking this challenge since GAs are very useful for extracting patterns in multiclass, high-dimensionality problems where heuristic knowledge is sparse or incomplete. The kernel parameters setting for SVM in a training process impacts on the classification accuracy. Feature selection is another factor that impacts classification accuracy. In this paper, the basic idea here consists in using a GA to optimize the parameters and to discover good subsets of genes simultaneously, without degrading the SVM classification accuracy. The GA feature extractor presented here utilizes feedback from the SVM classifier to the feature extractor Fig. 1 in section 5.

2 Discrete Cosine Transform

High information redundancy and correlation in face images result in inefficiencies when such images are used directly for recognition. DCT is a predominant tool first introduced by Ahmed et al. [7]. Since then, it was widely used as a feature extraction and compression in various applications on signal and image processing and analysis due to its fine properties, i.e., de-correlation, energy compaction, separability, symmetry and orthogonality. In face recognition, DCTs are used to reduce image information redundancy because only a subset of the transform coefficients are necessary to preserve the most important facial features.

3 SVM Classifier for Face Recognition

Recently, the SVM has been gaining popularity in the field of pattern classification due to its promising empirical performance, moderate computation complexity and its strong mathematical foundation. SVM are binary classifiers and different approaches like "one-against-all" and "one-against-one" are built to extend SVM to the multi-class classification case for face recognition [10]. The major method is the "one-against-one" method. This method constructs classifiers where each one is trained on data from two classes. For training data from the i^{th} and the j^{th} classes, we solve the following binary classification problem:

$$\min_{w^{ij}, b^{ij}, \xi^{ij}} \frac{1}{2} (w^{ij})^T w^{ij} + C \sum_t \xi_t^{ij} (w^{ij})^T$$

$$\begin{aligned}
& (w^{ij})^T \phi(x_t) + b^{ij} 1 - \xi_t^{ij}, \text{ if } y_t = i \\
& (w^{ij})^T \phi(x_t) + b^{ij} 1 - \xi_t^{ij}, \text{ if } y_t = i \\
& \xi_t^{ij} \geq 0
\end{aligned} \tag{1}$$

There are different methods for doing the future testing after all $p(p-1)/2$ classifiers are constructed. After some tests, we decide to use the following voting strategy suggested in [13]: if $\text{sign}((w^{ij})^T \phi(x_t) + b^{ij})$ says x is in the i^{th} class, then the vote for the i^{th} class is added by one. Otherwise, the j^{th} is increased by one. Then we predict x is in the class with the largest vote. The voting approach described above is also called the Max Wins strategy. In case that two classes have identical votes, thought it may not be a good strategy, now we simply select the one with the smaller index. Practically we solve the dual of (Eq. 1) whose number of variables is the same as the number of data in two classes. Hence if in average each class has l/k data points, we have to solve $k(k-1)/2$ quadratic programming problems where each of them has about $2l/k$ variables.

4 Genetic Algorithm for SVM Parameters Optimization

Goldberg [15] provides a nice introduction to GAs and the reader is referred to this source as well as the survey paper of Srinivas and Patnaik [8] for further information. The genetic algorithm is a method for solving both constrained and unconstrained optimization problems that is based on natural selection, the process that drives biological evolution. The genetic algorithm repeatedly modifies a population of individual solutions. At each step, the genetic algorithm selects individuals at random from the current population to be parents and uses them to produce the children for the next generation. Over successive generations, the population "evolves" toward an optimal solution. The genetic algorithm uses three main types of rules at each step to create the next generation from the current population: Selection, Crossover, and Mutation. In the literature, only a few algorithms have been proposed for SVM feature selection like in [9]. Some other GA-based feature selection methods were proposed [11][12]. However, these papers focused on feature selection and did not deal with parameters optimization for the SVM classifier. Therefore, in addition to the feature selection, proper parameters setting can improve the SVM classification accuracy. The choice of C and the kernel parameter is important to get a good classification rate. In the most case these parameters are tuned manually. In order to automatize this choice we use genetic algorithms. The SVM parameters, C and γ are real, we have to encode them with binary chains; we fix two search intervals, one for each parameter, $C_{\max} \leq C \leq C_{\min}$ and $\gamma_{\max} \leq \gamma \leq \gamma_{\min}$. To encode C and γ , we discretize the search spaces. Thus, a 32 bits encoding scheme of C is given by $C_b = C_{b1} \dots C_{b32}$ where:

$$C_b = \sum_{i=1}^{32} C_{bi} 2^{i-1} \tag{2}$$

and γ by $\gamma_b = \gamma_{b1} \dots \gamma_{b32}$ where:

$$\gamma_b = \sum_{i=1}^{32} \gamma_{bi} 2^{i-1} \tag{3}$$

with $C_b = g_{max}(C - C_{min})/(C_{max} - C_{min})$ and $\gamma_b = g_{max}(\gamma - \gamma_{min})/(\gamma_{max} - \gamma_{min})$ and $g_{max} = 2^{32} - 1$.

The fitness function used to evolve the chromosomes population is the SVM classification rate. The goal was to see if the GA would discover the work effectively. We lists some reasons why SVM must be used combined feature selection. One major weakness of SVMs is their high computational cost, which precludes real-time applications. In addition, SVMs are formulated as a quadratic programming problem and, therefore, it is difficult to use SVMs to do feature selection directly. Some researchers have proposed approximations to SVM for feature selection by first training the SVM using the whole training set, and then computing approximations to reduce the number of features.

5 Overview of the Proposed Method

The main steps of the proposed method are as follows:

1. FFS extraction using DCT.
2. Using Genetic Algorithms, in order to generate both the optimal feature subset and SVM parameters at the same time.
3. Classification of novel images.

Fig. 1 presents the general schema of feature selection and classification process. Firstly, a population of possible frequency features subset is genetically evolved, these features seems to be most useful to a particular classification problem from all those

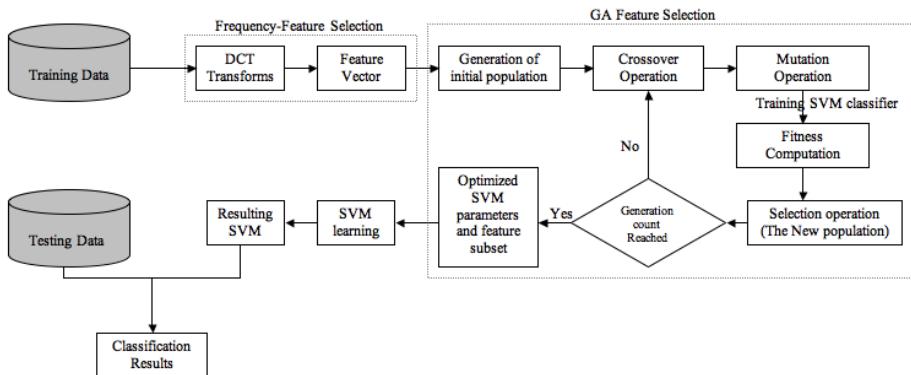


Fig. 1. The general process for gene subset selection and classification

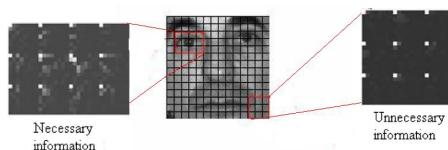


Fig. 2. Illustration of the effects of the block-based DCT for local appearance based face representation

available, it can be explain that they contain only highly informative and non-redundant features, which significantly improve classification. The genetic evolution is guided using the proposed fitness criterion, the quality of a given chromosome is proportional to the information gain measure computed using the dataset records retrieved from the training dataset, the chromosome comprises three parts, C, kernel parameter, and the features mask. The result is finally validated using a new test dataset. In fact, the basic idea here consists in using a GA to discover good subsets of genes, the goodness of a subset being evaluated by SVM classifier.

Using these methods we obtained three benefits, the first one that computational complexity is reduced as there is smaller number of inputs. Often, a secondary benefit found is that the accuracy of the classifier increases, and the last one is to remove the extra features (i.e like noise, obscuring other features from the learning algorithm) from a feature set, like unnecessary information showed in Fig. 2

6 The Dataset

To assess the robustness of our method against different facial expressions, lighting conditions and pose, we have collect grey-scale face images from two different face database available in the public domain, ORL face database¹ and Yale face database².



Fig. 3. Some samples from the used face database (ORL+Yale)

Face images selected are near frontal and contain variations in pose, illumination and expression. Eyebrows, eyes, nose, lips and surrounding area of face image contribute maximum in face recognition. So scale normalization of face images of data sets is carried out using the cropping phenomena which eliminate the unnecessary information from image and retain only internal structures. All the faces are then scaled to the size 48×48 pixels, aligned according to the eye positions. Sample images from the face databases are shown in Fig. 3. There are 330 subjects with 10 images per subject for a total of 3,300 images. The entire face database (ORL + Yale) is divided into two parts, six images of each subject are used to construct the training data and the remaining images are used for testing.

7 Experiment Results

We have performed a number of experiments in order to demonstrate the performance of the proposed approach on gray-scale images. In our study, the local information of

¹ http://www.cl.cam.ac.uk/Research/DTG/attarchive/pub/data/att_faces.zip

² <http://cvc.yale.edu/projects/yalefaces/yalefaces.html>

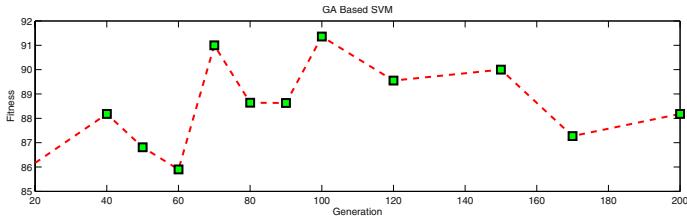


Fig. 4. The evolution of the best overall accuracy values for proposed approach

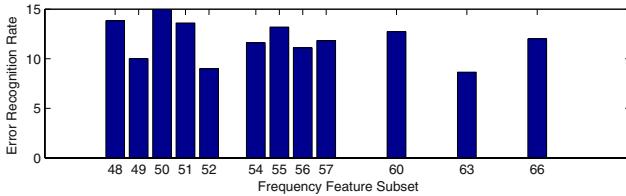


Fig. 5. Error Recognition Rate of SVM classifier as a function of the number of FFS selected by the GA approach

a candidate face can be obtained by using block-based DCT as follows: a face image is divided into blocks of 8 by 8 pixels size without overlapping. Each block is then represented by its DCT coefficients. From the obtained DCT coefficients only a small, *generic* feature set is retained in each block. Ekenel et al. [4] has proved that the highest information necessary to achieve high classification accuracy is contained in the first low frequency DCT coefficients via zigzag scanning.

The selection of features can have a considerable impact on the effectiveness of the resulting classification algorithm. Consider a feature set, $F = \{f_0, f_1, \dots, f_N\}$ which resulting from block-based DCT. The problem of feature selection can be seen as a case of feature weighting, where the numerical weights for each of the features have been replaced by binary values. A value of 1 could mean the inclusion of the corresponding feature into the subset, while a value of 0 could mean its absence. In a domain where objects are described by n features, there are 2^n possible feature subsets. Obviously, searching exhaustively for the best subset is futile. For this reason, the genetic algorithms has been identified as the best tools to explore such search space, and produce pseudo-optimal solutions that are sufficient to produce acceptable results. We run several experiments using the parameters for the designed GA as follows: population size: 50, crossover rate: 0.8, mutation rate: 0.01, and number of generations variate between 20 and 200.

As explained above, we used the GA approach to select a set of good FFS for SVMs classifier, the polynomial kernel has been found in our simulations to outperform linear and RBF kernel functions. In the present work, the library LIBSVM³ was used with a 5-fold cross-validation on the training data. We have computed histogram (see Figs. 5) showing the Error Recognition Rate as function of selected FFS.

³ <http://www.csie.ntu.edu.tw/~cjlin/libsvm>

In the AG, pairs of (C,d) are tried and the one with the best cross-validation accuracy is chosen. In the test set the highest accuracy was achieved with the SVM parameters pair (1, 1.5) (C, degree of polynomial kernel), which was obtained after several trials, while varying the dimensionality of the generation. The results are presented in Fig. 4. The performance improves immensely as the number of dimensions is increased from 60 to 70. Increasing the dimensionality from 70 to 100 provides only a relatively small improvement, while significantly increasing the amount of computation time required to generate the models. Based on this we have chosen 70 as the dimensionality of generation. The accuracy obtained by our algorithm during feature selection was 91%, using only 2.7% of the complete set of DCT coefficients under 70 generation. We can see in Fig. 4 that only 70 generation can be used to select FFS which contain a significant amount of person dependent information and got . Thus increasing the generation size to 200 implies a reduction in performance. This is verified in Fig. 5, where the vector length of FFS equal to 66 (generation size : 200) have worse performance than vector length of FFS equal to 52 (generation size : 70). We find that the 52 FFS has the maximal contribution value which is a criterion of feature importance to influence the Fitness. The pure DCT-SVM alone had an 87% recognition rate for the complete set of DCT coefficients.

8 Conclusions and Future Works

In conclusion, frequency feature subset based GA and SVM can further improve performance. An GA-SVM is proved to be effective in selecting FFS and significant fitness even if the sample set is very small. Thus, we have shown that genetic algorithms can play an important role in the automated loop of feature extraction and of classification. The obtained result shows that the genetic search can find very acceptable solutions for this problem in an acceptable run time, thereby obtaining a higher classification accuracy compared to DCT-SVM used only. In term of future works, this approach should be tested with other various datasets with different dimensions.

Acknowledgments

We gratefully acknowledge the support of the Maroc Telecom IAM (grant No. 105 909 05/PI) project 'Access Control' for this research. In addition, we thank Dr. Hicham LAANAYA for his contribution to the development of the GA feature extraction technique.

References

1. Chellappa, R., Wilson, C.L., Sirohey, S.: Human and machine recognition of faces: A survey. *Proceedings of the IEEE* 83(5), 705–740 (1995)
2. Koller, D., Sahami, M.: Towards optimal feature selection. In: ICML 1996, Bari, Italy, pp. 87–95 (1996)
3. Brill, F., Brown, D., Martin, W.: Fast genetic selection of features for neural network classifiers. *IEEE Transaction on Neural Networks* 3(2), 324–328 (1992)

4. Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., Vapnik, V.: Feature selection for support vector machines. *Advances in Neural Information Processing Systems* (2000)
5. Jain, A.K., Zongker, D.: Feature selection: Evaluation, application, and small sample performance. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 19(2), 153–158 (1997)
6. Devijverand, P.A., Kittler, J.: *PatternRecognition:AStatisticalApproach*. Prentice Hall, London (1982)
7. Ahmed, N., Natarajan, T., Rao, K.R.: Discrete Cosine Transform. *IEEE Transactions Computers* 23, 90–94 (1974)
8. Srinivas, M., Patnaik, L.: Genetic algorithms: A survey. *IEEE Computer* 27, 1726 (1994)
9. Bradley, P.S., Mangasarian, O.L.: Feature selection via concave minimization and support vector machines. In: *Proceedings of the 13th international conference on machine learning*, San Francisco, CA, pp. 8290 (1998)
10. Burges, J.C.: A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery* 2(2), 121–167 (1998)
11. Raymer, M.L., Punch, W.F., Goodman, E.D., Kuhn, L.A., Jain, A.K.: Dimensionality reduction using genetic algorithms. *IEEE Transactions on Evolutionary Computation* 4(2), 16417 (2000)
12. Yang, J., Honavar, V.: Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems* 13(2), 4449 (1998)
13. Friedman, J.: Another approach to polychotomous classification Technical report, Department of Statistics, Stanford University (1996)
14. Ekenel, H.K., Stiefelhagen, R.: Local Appearance based Face Recognition Using Discrete Cosine Transform. In: *EUSIPCO 2005*, Antalya, Turkey, vol. 23(7) (2005)
15. Goldberg, D.: *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading (1989)

A Path Following Algorithm for Graph Matching

Mikhail Zaslavskiy^{1,3,4}, Francis Bach², and Jean-Philippe Vert^{1,4}

¹ The Center for Computational Biology, École des Mines de Paris, ParisTech, 35 rue Saint-Honoré, Fontainebleau, France

² INRIA-Willow Project, École Normale Supérieure, 45 rue d'Ulm, Paris, France

³ The Center for Mathematical Morphology, École des Mines de Paris, ParisTech, 35 rue Saint-Honoré, Fontainebleau, France

⁴ Institut Curie, Section Recherche, INSERM U900, Paris, France

Abstract. We propose a convex-concave programming approach for the labelled weighted graph matching problem. The convex-concave programming formulation is obtained by rewriting the graph matching problem as a least-square problem on the set of permutation matrices and relaxing it to two different optimization problems: a quadratic convex and a quadratic concave optimization problem on the set of doubly stochastic matrices. The concave relaxation has the same global minimum as the initial graph matching problem, but the search for its global minimum is also a complex combinatorial problem. We therefore construct an approximation of the concave problem solution by following a solution path of the convex-concave problem obtained by linear interpolation of the convex and concave formulations, starting from the convex relaxation. The algorithm is compared with some of the best performing graph matching methods on three datasets: simulated graphs, QAPLib and handwritten Chinese characters.

1 Introduction

The graph matching problem consists in finding a correspondence between vertices of two given graphs which is optimal in a certain sense¹. This problem plays a very important role in various fields of pattern recognition like optical character recognition [1], image analysis (2D and 3D) [23], or bioinformatics [4].

Because of the combinatorial nature of the problem, it is very hard to solve it exactly for large graphs; however some methods based on incomplete enumeration may be applied to search for an exact optimal solution in the case of small or sparse graphs. An alternative is to use approximate algorithms which are supposed to be much more scalable, but only find an approximate solution. This group is represented by two classes of methods: spectral methods [5,6,2,7] and methods which work directly with graph adjacency matrices and typically involve a relaxation of the discrete optimization problem [8,9,10,3].

An interesting instance of the graph matching problem is the matching of labeled graphs. If there is also a similarity measure between labels, then we

¹ Usually, a correspondence is optimal if it aligns graph structures in the best way, but other criteria are possible as well.

search a correspondence which matches not only the structures of the graphs but also vertices with similar labels. Some widely used approaches only use the information about similarities between graph labels [11], some try to combine information on graph structures and vertex similarities [10, 4]. We do not discuss here the various graph matching algorithms, a good review of graph matching algorithms may be found in [12].

In this article, we propose an approximate methods for labeled weighted graph matching. This method is based on a formulation of the labeled graph matching problem as a quadratic assignment problem (QAP) over the set of permutation matrices, where the quadratic term encodes the structural compatibility and an additional linear term encodes vertex compatibilities. We propose two relaxations of this problem, resulting in one quadratic convex and one quadratic concave optimization problem on the set of doubly stochastic matrices. While the concave relaxation has the same global minimum as the initial QAP, it also does not have an efficient optimization algorithm. We find a local minimum of this problem by following a solution path of a family of convex-concave optimization problems, obtained by linearly interpolating the convex and concave relaxations. Starting from the convex formulation with a unique local (and global) minimum, the solution path leads to a local optimum of the concave relaxation. We perform an extensive comparison of this **PATH** algorithm with several state-of-the-art matching methods on small simulated graphs and various QAP benchmarks, and show that it consistently provides state-of-the-art performances while scaling to graphs of up to a few thousands vertices for a modern computer. We further illustrate the use of the algorithm on the matching of handwritten Chinese characters. Additional simulations may be found in [13].

2 Problem Description

A graph $G = (V, E)$ of size N is defined by a finite set of vertices $V = \{1, \dots, N\}$ and a set of edges $E \subset V \times V$. We consider weighted undirected graphs with no self-loop, i.e., all edges (i, j) have an associated positive real value $w(i, j) = w(j, i)$ and $w(i, i) = 0 \forall i, j \in V$. Each such graph can be equivalently represented by a symmetric adjacency matrix A where $A_{ij} = w(i, j)$.

Given two graphs G and H with the same number of vertices N^2 , the problem of matching G and H consists in finding a correspondence between vertices of G and vertices of H which aligns G and H in some optimal way. The correspondence between vertices may be defined by a permutation matrix P , P_{ij} is equal to 1 if the i -th vertex of G is matched to the j -th vertex of H , and 0 otherwise. After applying the permutation defined by P to the vertices of H we obtain a new graph isomorphic to H which we denote by $P(H)$. The adjacency matrix of the permuted graph, $A_{P(H)}$, is simply obtained from A_H by the equality $A_{P(H)} = P A_H P^T$.

In order to assess whether a permutation P defines a good matching between the vertices of G and those of H , a quality criterion must be defined. We focus

² Otherwise the smallest may be completed with dummy nodes.

in this paper on measuring the discrepancy between the graphs after matching of edges which are present in one graph and not in the other one:

$$F(P) = \|A_G - A_{P(H)}\|_F^2 = \|A_G - PA_H P^T\|_F^2, \quad (1)$$

where $\|\cdot\|_F$ is the Frobenius matrix norm. Therefore, the problem of graph matching can be reformulated as the problem of minimization of $F(P)$ over the set of permutation matrices.

An interesting generalization of the graph matching problem is the problem of labeled graph matching. Here each graph has associated labels to all its vertices and the objective is to find an alignment that fits well graph labels and graph structures at the same time. If we let C_{ij} denote the cost of fitness between i -th vertex of G and j -th vertex of H then the matching problem based only on label comparison can be formulated as follows

$$\min_P \text{tr}(C^T P) = \sum_{i=1}^N \sum_{j=1}^N C_{ij} P_{ij} = \sum_{i=1}^N C_{i,P(i)}. \quad (2)$$

A natural way of unifying of (2) and (1) is a linear combination

$$\min_P \{(1-\alpha)F(P) + \alpha \text{tr}(C^T P)\}. \quad (3)$$

In the next section we describe our new algorithm which is based on the technique of convex-concave relaxation of the initial problem (1).

3 Convex-Concave Relaxation

The criterion of graph matching problem we consider is (1). Since permutation matrices are also orthogonal matrices, we can rewrite $F(P)$ on \mathcal{P} as follows:

$$F_0(P) = \|A_G - PA_H P^T\|_F^2 = \|(A_G - PA_H P^T)P\|_F^2 = \|A_G P - PA_H\|_F^2. \quad (4)$$

The graph matching problem is then the problem of minimizing $F_0(P)$ over \mathcal{P} , which we call **GM**:

$$\mathbf{GM}: \min_{P \in \mathcal{P}} F_0(P). \quad (5)$$

A first relaxation of **GM** is obtained by expanding the convex quadratic function $F_0(P)$ on the set of doubly stochastic matrices \mathcal{D} :

$$\mathbf{QCV}: \min_{P \in \mathcal{D}} F_0(P). \quad (6)$$

The **QCV** problem may be solved in polynomial time (this is a convex quadratic minimization problem) by the Frank-Wolfe algorithm [4], but the optimal value usually does not belong to the set of the extreme points of \mathcal{D} so we have to use the approximation: $\arg \min_{P \in \mathcal{P}} F(P) \approx \Pi_{\mathcal{P}} \arg \min_{D \in \mathcal{D}} F(P)$. The projection $\Pi_{\mathcal{P}}$ may be made by the Hungarian algorithm in $O(N^3)$, but the problem is that if $\arg \min_{D \in \mathcal{D}} F(P)$ is far from \mathcal{P} then the approximation quality may be poor.

We now present a second relaxation of **GM**, resulting in a concave minimization problem. For that purpose, let us introduce the diagonal degree matrix D of an adjacency matrix A , which is the diagonal matrix with entries $D_{ii} = d(i) = \sum_{j=1}^N A_{ij}$ for $i = 1, \dots, N$, as well as the Laplacian matrix $L = D - A$. A having only nonnegative entries, it is well-known that the Laplacian matrix is positive semidefinite [15]. We can now rewrite $F(P)$ as follows:

$$\begin{aligned} F_0(P) &= \|A_G P - PA_H\|_F^2 = \|(D_G P - PD_H) - (L_G P - PL_H)\|_F^2 \\ &= \|D_G P - PD_H\|_F^2 - 2\text{tr}[(D_G P - PD_H)^T (L_G P - PL_H)] + \|L_G P - PL_H\|_F^2. \end{aligned}$$

After some transformations $F_0(P)$ may be rewritten as

$$F_0(P) = -\text{tr}(\Delta P) + \text{tr}(L_G^2) + \text{tr}(L_H^2) - 2\text{vec}(P)^T (L_H \otimes L_G) \text{vec}(P), \quad (7)$$

where we introduced the matrix $\Delta_{i,j} = (D_H(j,j) - D_G(i,i))^2$ and we used \otimes to denote the Kronecker product of two matrices. Let denote $F_1(P)$ the function defined in [7] expanded on \mathcal{D} . Since graph Laplacian matrices are positive semidefinite, the matrix $L_H \otimes L_G$ is also positive semidefinite as a Kronecker product of two symmetric positive semi-definite matrices [16]. Therefore the function $F_1(P)$ is concave on \mathcal{D} , and we obtain a concave relaxation of the graph matching problem:

$$\mathbf{QCC}: \min_{P \in \mathcal{D}} F_1(P). \quad (8)$$

Interestingly, the global minimum of a concave function is necessarily at one of the extreme points of the convex set where it is minimized, so the minimum of $F_1(P)$ on \mathcal{D} is in fact in \mathcal{P} . At this point, we have obtained two relaxations of **GM** as optimization problems on \mathcal{D} : the first one is the convex minimization problem **QCV** [6], which can be solved efficiently but leads to a solution in \mathcal{D} that must then be projected onto \mathcal{P} , and the other is the concave minimization problem **QCC** [8] which can not be solved efficiently but has the same solution as the initial combinatorial problem **GM**.

We propose to approximately solve **QCC** by tracking a path of local minima over \mathcal{D} of a sequence of functions that linearly interpolate between $F_0(P) := F(P)$ and $F_1(P)$, namely:

$$F_\lambda(P) = (1 - \lambda)F_0(P) + \lambda F_1(P),$$

for $0 \leq \lambda \leq 1$. For all $\lambda \in [0, 1]$, F_λ is a quadratic function. We recover the convex function F_0 for $\lambda = 0$, and the concave function F_1 for $\lambda = 1$. Our method searches sequentially local minima of F_λ , where λ moves from 0 to 1. More precisely, we start at $\lambda = 0$, and find the unique local minimum of F_0 (which is in this case its unique global minimum) by any classical QP solver. Then, iteratively, we find a local minimum of $F_{\lambda+d\lambda}$ given a local minimum of F_λ by performing a local optimization of $F_{\lambda+d\lambda}$ starting from the local minimum of F_λ , using for example the Frank-Wolfe algorithm. Repeating this iterative process for $d\lambda$ small enough we obtain a path of solutions $P^*(\lambda)$, where $P^*(0) = \arg \min_{P \in \mathcal{D}} F_0(P)$ and $P^*(\lambda)$ is a local minimum of F_λ for all $\lambda \in [0, 1]$. Noting

that any local minimum of the convex function F_1 on \mathcal{D} is in \mathcal{P} , we finally output $P^*(1) \in \mathcal{P}$ as an approximate solution of **GM**.

The pseudo-code for this **PATH** algorithm is presented below.

```

1. Initialization:
  (a)  $\lambda := 0$ 
  (b)  $P^*(0) = \arg \min F_0$  — standard QP solver.
2. Cycle over  $\lambda$ :
   do
     (a)  $\lambda_{new} := \lambda + d\lambda$ 
     (b) if  $|F_{\lambda_{new}}(P^*(\lambda)) - F_{\lambda_{new}}(P^*(\lambda))| < \epsilon$  then  $\lambda = \lambda_{new}$ 
        else
           $P^*(\lambda_{new}) = \arg \min F_{\lambda_{new}}$  is found by Frank-Wolfe starting from  $P^*(\lambda)$ 
           $\lambda = \lambda_{new}$ 
        endif
   while  $\lambda < 1$ 
3. Output:  $P^{out} := P^*(1)$ 
```

Of course, in spite of these justifications the **PATH** algorithm only gives an approximation of the global minimum in the general case. For a more detailed analysis, see [13].

If we match two labeled graphs, then we may increase the performance of our method by using information on pairwise similarities between their nodes:

$$\min_P F_\lambda^\alpha(P) = \min_P (1 - \alpha)F_\lambda(P) + \alpha \text{tr}(C^T P). \quad (9)$$

The advantage of the last formulation is that $F_\lambda^\alpha(P)$ is just $F_\lambda(P)$ with an additional linear term. Therefore we can use the same algorithm for $F_\lambda^\alpha(P)$ optimization that we have used before for $F_\lambda(P)$.

4 Results

Synthetic examples. In this section we compare the proposed algorithm with some classical methods on the example of artificially generated graphs. Our choice of random graph types is based on [17]. Each type of random graphs is defined by the distribution function of node degree, we consider three distributions: binomial (bin), exponential (exp) and power law (pow). If we are interested in isomorphic graph matching then we compare just the initial graph and its randomly permuted copy. To test the matching of non-isomorphic graphs, we add randomly $\sigma_n N_E$ edges to the initial graph and to its permuted copy, where N_E is the number of edges in the original graph, and σ_n is the noise level.

The first series of experiments are experiments on small size graphs ($N=8$), here we are interested in comparison of **PATH** algorithm, **QCV** approach [6], Umeyama spectral algorithm "U" [5], a linear programming approach **LP** [8] and exhaustive search **OPT** which is feasible for the small size graphs. The results are presented on the top line in Figure 1. The second row of Figure 1 presents

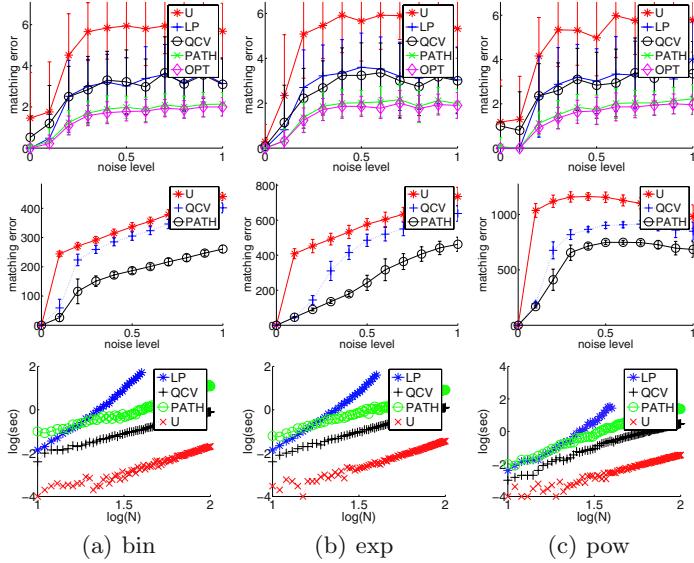


Fig. 1. Precision and timing of graph matching algorithms

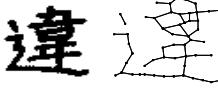
results of experiments on large graphs ($N=100$). In all cases **PATH** algorithm works much better than all other algorithms. Note, that the application of solution path is useful (**PATH** algorithms works much better than **QCV**). A very remarkable thing is that **PATH** solution is very close to the global minimum. The third row presents the algorithm complexity (i.e the time of computing time as a function of graph size), **U** algorithm is known to be one of the fastest graph matching algorithm, **PATH** and **QCV** show the same order of complexity (slope of the corresponding curves ≈ 3), complexity of **LP** is $O(N^7)$. All results are coherent with theoretical values of algorithm complexities.

QAP benchmark library. The problem of graph matching may be considered as a particular case of the quadratic assignment problem (QAP). [10] proposed the **QPB** graph matching algorithm and tested it on matrices from the QAP benchmark library (see [18]), as well as graduated assignment algorithm **GRAD** [9] and Umeyama. Results of **PATH** application to the same matrices are presented in the table (II). **PATH** outperforms **QPB**, **GRAD** and **U** almost in all cases.

Recognition of handwritten chinese characters. Another example that we consider in this paper is the problem of handwritten chinese character recognition from the ETL9B database. We use a score of optimal matching as a similarity measure in KNN for character classification. Here we compare the performance of four methods: linear support vector machine (SVM), SVM with gaussian kernel, KNN based on score of shape context matching and K-nearest neighbor (KNN) based on score of graph matching. As score we use just the value of the objective

Table 1. QAP benchmark library

QAP	MIN	PATH	QPB	GRAD	U
chr12c	11156	18048	20306	19014	40370
chr15a	9896	19086	26132	30370	60986
chr15c	9504	16206	29862	23686	76318
chr20b	2298	5560	6674	6290	10022
chr22b	6194	8500	9942	9658	13118
esc16b	292	300	296	298	306
rou12	235528	256320	278834	273438	295752
rou15	354210	391270	381016	457908	480352
rou20	725522	778284	804676	840120	905246
tai10a	135028	152534	165364	168096	189852
tai15a	388214	419224	455778	451164	483596
tai17a	491812	530978	550852	589814	620964
tai20a	703482	753712	799790	871480	915144
tai30a	1818146	1903872	1996442	2077958	2213846
tai35a	2422002	2555110	2720986	2803456	2925390
tai40a	3139370	3281830	3529402	3668044	3727478

character 1	character 2	character 3
		

Method	CV	STD
Linear SVM	0.377	± 0.090
SVM with gaussian kernel	0.359	± 0.076
KNN ($\lambda=1$): shape context	0.399	± 0.081
KNN ($\lambda=0.4$)	0.248	± 0.075
KNN ($\lambda=0$): pure graph matching	0.607	± 0.072

Fig. 2. Chinese characters.(*CV*, *STD*)—mean and standard deviation of test error over cross-validation runs (five folds, 50 repetition)

function (9) in the optimal point. We have selected three chinese characters which are the most difficult to distinguish. Examples of these characters as well as examples of extracted graphs are presented in Figure 2. In SVM based algorithms we use directly the values of image pixels (so each image is represented by a binary vector), in graph matching algorithm we use binary adjacency matrices of extracted graphs and shape context matrices (see 11). Our data set consist of 50 exemplars (images) of each class. To compare different methods we use the cross validation error (five folds). Complete results may be found in Figure 2.

5 Conclusion

We have presented the **PATH** algorithm, a new method for graph matching based on convex-concave relaxations of the initial integer programming problem. It obtained very promising results in all experiments against state-of-the-art methods, and has the same empirical complexity as the fastest graph matching algorithms. It furthermore allows to easily integrate the information on label similarities. In future works it will be interesting to investigate other choices of convex and concave relaxations, and to generalize the approach to directed graphs, i.e., asymmetric adjacency matrices, which could provide new methods for the general QAP.

References

1. Filatov, A., Gitis, A., Kil, I.: Graph-based handwritten digit string recognition. In: Third International Conference on Document Analysis and Recognition (ICDAR 1995), pp. 845–848 (1995)
2. Carcassoni, M., Hancock, R.E.: Spectral correspondance for point pattern matching. *Pattern Recognition* 36, 193–204 (2002)
3. Schellewald, C., Schnor, C.: Probabilistic subgraph matching based on convex relaxation. Lecture notes in computer science, pp. 171–186. Springer, Heidelberg (2005)
4. Singh, R., Xu, J., Berger, B.: Pairwise global alignment of protein interaction networks by matching neighborhood topology. *Research in Computational Molecular Biology* 4453, 16–31 (2007)
5. Umeyama, S.: An eigendecomposition approach to weighted graph matching problems. *Transaction on pattern analysis and machine intelligence* 10 (1988)
6. Shapiro, L., Brady, J.: Feature-based correspondance: an eigenvector approach. *Image and vision computing* 10, 283–288 (1992)
7. Caelli, T., Kosinov, S.: An eigenspace projection clustering method for inexact graph matching. *TPAMI* 24 (2004)
8. Almohamad, H., Duffuaa, S.O.: A linear programming approach for the weighted graph matching problem. *TPAMI* 15 (1993)
9. Gold, S., Rangarajan, A.: A graduated assignment algorithm for graph matching. *Transaction on pattern analysis and machine intelligence* 18 (1996)
10. Cremers, D., Kohlberger, T., Schnor, C.: Evaluation of convex optimization techniques for the graph-matching problem in computer vision. *Patter Recogn.* 2191 (2001)
11. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. *Transaction on pattern analysis and machine intelligence* 24 (2002)
12. Conte, D., Foggia, P., Sansone, C., Vento, M.: Thirty years of graph matching in pattern recognition. *International journal of pattern recognition and AI* 18, 265–298 (2004)
13. Zaslavskiy, M., Bach, F., Vert, J.P.: Path following algorithm for graph matching problem. arXiv:0801.3654v1 (2008)
14. Frank, M., Wolfe, P.: An algorithm for quadratic programming. *Naval Research Logistics Quarterly* 3, 95–110 (1956)

15. Chung, F.R.K.: Spectral Graph Theory. Americal Mathematical Society (1997)
16. Golub, G.H., Loan, C.F.V.: Matrix computations, 3rd edn. Johns Hopkins University Press, Baltimore (1996)
17. Newman, M.E.J., Strogatz, S.H., Watts, D.J.: Random graphs with arbitrary degree distributions and their applications. PHYSICAL REVIEW 64 (2001)
18. Cela, E.: Qaudratuc assignment problem library (2007), <http://www.opt.math.tu-graz.ac.at/qplib/>

Statistically Valid Graph Representations of Scale-Space Geometry

Tomoya Sakai and Atsushi Imiya

Institute of Media and Information Technology, Chiba University, Japan
`{tsakai,imiya}@faculty.chiba-u.jp`

Abstract. This paper presents a statistical scale-selection criterion for graph representations derived from differential geometric features of a greyscale image in a Gaussian scale space. The image gradient in scale space derives hierarchical and topological relationships among the bright and dark components in the image. These relationships can be represented as a tree and a skeleton-like graph, respectively. Since the image at small scales contains invalid geometric features due to noise and numerical errors, a validation scheme is required for the detected features. The presented scale-selection criterion allows us to identify the valid features used for the graph representations with statistical confidence.

1 Introduction

We propose graph representations of an greyscale image and their statistical validation scheme wholly integrated in a Gaussian scale-space framework. The scale-space theory [1][2][3][4] is a mathematical framework for multiscale image representation, which clarifies hierarchical relationships and topological relationships among geometric features of the image such as critical points and isophotes [5][6][9][10][11], and separatrices [7][8]. Graph representations of these relationships among the image features are available for abstracting the structure of the image.

In general, the image analyses using a pre-smoothing filter suffer from the selection of the filter bandwidth. This problem can be interpreted as selection of the scale at which invalid features of the image due to noise are suppressed without smoothing away the valid image features for the analyses. Therefore, we require a scheme to validate the image features.

In this paper, we present a scale-selection criterion for two graph representations of an image: (i) scale-space tree for the hierarchical structure of the image and (ii) scale-space skeleton for the topological structure of the image. These graph representations are derived from the image gradient in scale space. Our scale-selection criterion is based on a statistical approach similar to scale-based clustering methods [12][13][14][15].

2 Scale Space Basics

2.1 Gaussian Scale Space

We define the image as a nonnegative scalar function of one or more variables whose domain is the extended real space.

Definition 1. An image is defined as a nonnegative scalar function $f(\mathbf{x})$, $\mathbf{x} \in \overline{\mathbb{R}}^d$.

Here, $\overline{\mathbb{R}}^d$ denotes d -dimensional extended real space, which includes a point at infinity. Although the domain of a greyscale image is practically bounded within a limited area or volume, we embed such an image in the extended real space to apply our scale-space theory.

The scale space is known as a set of blurred images derived from conversion of a scale or an inverse resolution of the original image $f(\mathbf{x})$. A one-parameter family of positive function $f(\mathbf{x}, \tau)$ is obtained from the blurring filter with a single control parameter τ . Axiomatic derivation of the Gaussian kernel for the scale-space filtering was proposed by Witkin [1] and Koenderink [2].

$$f(\mathbf{x}, \tau) = G(\mathbf{x}, \tau) * f(\mathbf{x}) \quad (1)$$

Here, $G(\mathbf{x}, \tau)$ is an isotropic Gaussian function

$$G(\mathbf{x}, \tau) = \frac{1}{\sqrt{4\pi\tau}^d} \exp\left(-\frac{|\mathbf{x}|^2}{4\tau}\right). \quad (2)$$

The Gaussian kernel in Eq. (2) is normalised so that the scale-space image $f(\mathbf{x}, \tau)$ satisfies an isotropic linear diffusion equation

$$\frac{\partial}{\partial\tau}f(\mathbf{x}, \tau) = \Delta f(\mathbf{x}, \tau), \quad f(\mathbf{x}, 0) = f(\mathbf{x}). \quad (3)$$

In this study, we deal with the Gaussian scale space in the extended real scale and space. The point at infinity and the infinite scale play essential roles in the structural analysis of the image in scale space.

Definition 2. The Gaussian scale-space image $f(\mathbf{x}, \tau)$, $(\mathbf{x}, \tau) \in (\overline{\mathbb{R}}^d, \overline{\mathbb{R}}^+)$ is the convolution of the d -dimensional original image $f(\mathbf{x}) \geq 0$ with the isotropic Gaussian kernel $G(\mathbf{x}, \tau)$.

The Gaussian scale-space image $f(\mathbf{x}, \tau)$ is a diffused version of the original image $f(\mathbf{x})$ in the infinite domain. As the features of the original image are reduced with increasing scale, the structure of the image is simplified.

2.2 Differential Geometric Features of Scale Space Image

Critical Points Critical points are well-defined feature points of the scale-space image $f(\mathbf{x}, \tau)$ in the sense of the differential geometry.

Definition 3. *The critical points are defined as points where the spatial gradient vanishes.*

$$\{(\mathbf{x}, \tau) \mid \nabla f(\mathbf{x}, \tau) = \mathbf{0}\} \quad (4)$$

A regular critical point has a nonzero determinant of the Hessian matrix $\mathbf{H} = \nabla \nabla^\top f(\mathbf{x}, \tau)$. The regular critical points can be classified into $d + 1$ types based on the combination of signs of the eigenvalues of \mathbf{H} . We denote the set of signs of the eigenvalues as (\pm, \pm, \dots, \pm) . For example, if $d = 2$ we have three types of critical points: local maximum $(-, -)$, saddle $(+, -)$, and local minimum $(+, +)$.

The critical points can be considered as geometric cues. The local maxima are representative of bright image components. The local minima correspond to dark cavities in the image. Note that the point at infinity is a hidden local minimum for a positive image. The local minimum at infinity represents the dark background of the positive image. The saddle points reside between the local extrema, and imply connections between them.

Critical Curves Observing these critical points in the scale space, we see their trajectories across scales called the critical curves.

Definition 4. *The critical curves are the trajectories of critical points in scale space.*

Figure 1(a) shows an example of the critical curves in scale space. The critical curves are also classified as local maximum curves, saddle curves, and local minimum curves. Since a saddle point exists between local extrema, disappearance of a local extremum with increasing scale is always accompanied by that of the saddle point. The local extrema are annihilated when they meet the saddle points with increasing scale. The annihilation point is a singular point where at least one of the eigenvalues of \mathbf{H} is zero. Therefore, a local extremum curve and a saddle curve can share the singular point in the scale space. The connections of the critical curves in scale space indicate topological relationships among the local extrema and saddle points.

Figure Field. In thermal physics, the diffusion equation describes transmission of heat. The solution and its spatial gradient are the transition of temperature distribution and the flow of heat. In the scale space of an image, $f(\mathbf{x}, \tau)$ and its spatial gradient are analogically the transition of the distribution of image intensity and the *flow of image intensity*. Therefore, the spatial gradient field of the scale-space image $f(\mathbf{x}, \tau)$ and its field curves are called the figure field and the figure-flow curves, respectively [5].

Definition 5. *The figure field \mathbf{F} in the isotropic space at a scale τ is defined as the negative of the gradient vector field of the scale-space image.*

$$\mathbf{F} = -\nabla f(\mathbf{x}, \tau) \quad (5)$$

Definition 6. *The figure-flow curves are the directional lines of the figure field.*

Figure 1(b) illustrates an example of the figure-flow curves of a two-dimensional scale-space image at a fixed scale. The local maxima, local minima, and saddle points are the sources, drains, and confluent points of the flow of image intensity.

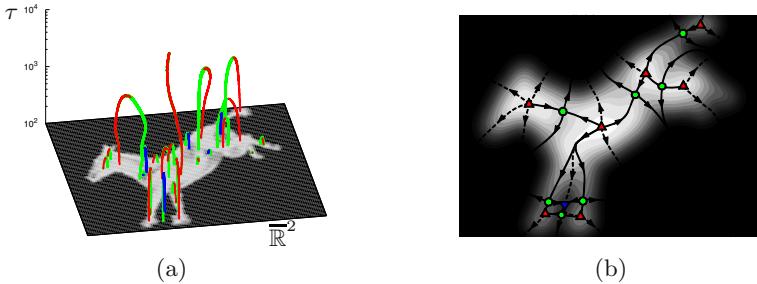


Fig. 1. Critical curves and figure-flow curves. (a) Critical curves in scale space for an image of a kicking horse [16]. (b) Figure-flow curves of the image at $\tau = 400$. The solid curve indicates the separatrix. The up triangle, down triangle, and circle indicate the local maximum, local minimum, and saddle point, respectively.

Separatrices. The topological relationships among the critical points can be analysed by the figure-flow curves. In two-dimensional case, a saddle $(+, -)$ has a pair of each inward and outward flow curves called the *separatrices*. A separatrix of the inward flow curves connects between local maxima of $f(\mathbf{x}, \tau)$, i.e., the sources of the flow.

3 Graph Representations of Scale-Space Geometry

3.1 Scale-Space Tree

The critical curves in scale space provides a topological relationships between the critical points. It is suggested that the singular point shared by a pair of critical curves is connected to another local extremum by the so-called *antidirectional figure-flow curve* at the scale of annihilation as shown in Fig. 2(a) [11]. Since the figure-flow curve is the path of the flow of image intensity, the antidirectional figure-flow curve indicates the source or drain of the flow to or from the singular point. Consequently, hierarchical relationships among the critical points are exhibited in the scale space by the critical curves and the antidirectional figure-flow curves.

The hierarchical relationships are symbolically expressed as a tree. We present an algorithm for bottom-up construction of a tree T .

Algorithm I: Scale-Space Tree

1. Set nodes for critical points at an initial scale $\tau = \tau_0$ to be the leaves of T .
2. Diffuse the image $f(\mathbf{x}, \tau_{n-1})$, yielding $f(\mathbf{x}, \tau_n)$.
3. Detect annihilation events of local extrema and saddles within $\tau_{n-1} < \tau \leq \tau_n$ by examining links of the critical points between the subsequent scales τ_{n-1} and τ_n .
4. For each detected annihilation event:
 - i Add to T a new node with two branches leading to nodes for the local extremum and saddle involved in annihilation.

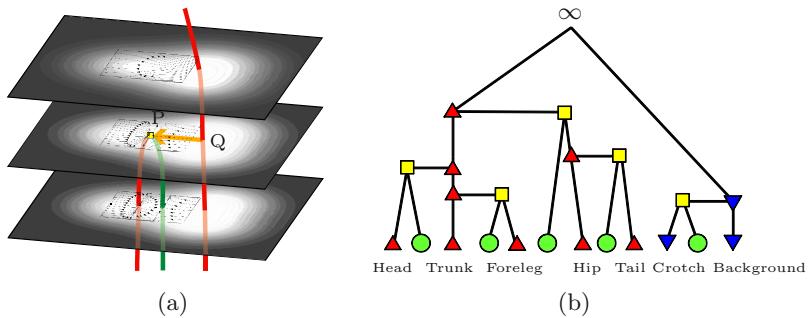


Fig. 2. Graph representation of hierarchical relationships among the critical points. (a) The singular point P is connected to the local maximum Q by the antidiirectional figure-flow curve, which indicates subordinate relationship between them. (b) A few levels of scale-space tree of the kicking horse image in Fig. 1. The square indicates the singular point. ∞ at the root indicates the virtual annihilation.

- ii For the annihilation event of the local maximum, maximise $f(\mathbf{x}, \tau_n)$ using the annihilation point for the initial point.
- iii For the annihilation event of the local minimum, minimise $f(\mathbf{x}, \tau_n)$ using the annihilation point for the initial point. Regard the annihilation point as having a connection to a local minimum at infinity if minimisation results in a point on the image boundary.
- iv Add to T a new node with two branches. One branch is attached to the node for annihilation point. The other branch is attached to a node for the local extremum determined by the maximisation or minimisation.
- 5. Return to step 2 until there remains no saddle point.
- 6. Add to T a node for the virtual annihilation point of one remaining local maximum and the local minimum at infinity.

An example of the scale-space tree is shown in Fig. 2(b).

We store in the nodes of T the scales of annihilation. The scales of annihilation can be used for a statistical test of the validity of the critical points, which is discussed in Section 4. Note that Algorithm I is not designed to detect the creation event [7910] because we are interested in the hierarchical relationships among the critical point at the initial scale τ_0 .

3.2 Scale-Space Skeleton

We present a novel graph representation of topological relationships among local maxima of the image. Since the local maxima are the sources of the flow of image intensity with respect to scale, the critical curves of local maxima and the antidiirectional figure-flow curves indicate how the distributed image intensity is inherited from coarse to fine. At the coarsest scale, one remaining local maximum represents the total image intensity. As the scale decreases, the local maxima appear when the image intensity is supplied by the existing local

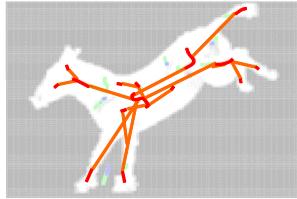


Fig. 3. Scale-space skeleton of the kicking horse

maxima through the antidiirectional figure-flow curves. The inheritance of the image intensity is symbolically expressed by a tree with nodes related to the local maxima embedded in the scale-space tree.

Geometrically, the critical curves $\{(\mathbf{x}, \tau) | \nabla f(\mathbf{x}, \tau) = \mathbf{0}, \mathbf{x} \in \overline{\mathbb{R}}^d, \tau \geq \tau_0\}$ in scale space can be projected onto the space $\overline{\mathbb{R}}^d$ as one-dimensional manifolds $\{\mathbf{x}(\tau) | \mathbf{x}(\tau) \in \overline{\mathbb{R}}^d, \tau \geq \tau_0\}$. Here, τ_0 is an arbitrary scale. The antidiagonal figure-flow curves at $\tau \geq \tau_0$ can also be projected onto $\overline{\mathbb{R}}^d$ as the curves, which supply the linkage between the manifolds $\{\mathbf{x}(\tau)\}$. We call the resulting graph the scale-space skeleton, which exhibits trajectories of the representative points of the flow of image intensity. The nodes of the graph represent critical points at the scale τ_0 , and all singular points and their source points. The edges of the graph are the projected critical curves and antidiagonal figure-flow curves. Figure 3 is an illustration of the scale-space skeleton. A skeleton-like representation of the image is obtained from the scale space.

4 Statistical Selection of Scale

In general, most of the image analyses using Gaussian filtering suffer from the selection of the filter bandwidth, i.e., the scale. The image loses its geometric features at large scales. The detected features, however, are invalid at small scales because the geometric features are infected by unfiltered noise. The differential operations enhance the noise, especially. As the result, the detected features are so random and experimentally less reproducible. We require a validation scheme to identify the image features with the statistically significant reproducibility.

Clustering methods of data points using scale space [12][13][14] give us a hint of the validation of the geometric features of image. In these clustering methods, the data clusters are related to the modes of the probability density function (PDF) of the data points. The PDF is estimated by taking the sum of the Gaussian kernel functions at the data points, which satisfies the scale-space axioms. The cluster validity is attributed to that of the estimation of modes, or the local maxima of the estimated PDF. We have proposed a statistical criterion to identify the valid clusters by the *life* of mode [15]. The life is defined as

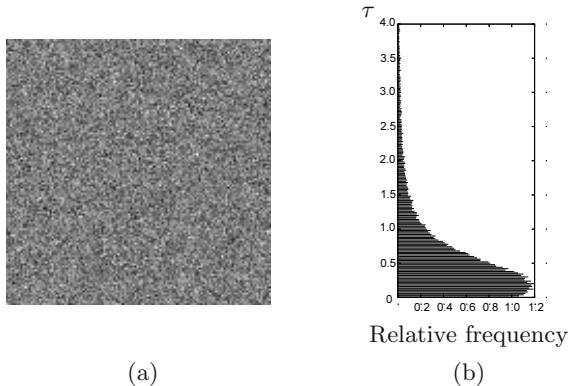


Fig. 4. Noise image and life histogram for local maxima. (a) The noise image has uncorrelated random pixel values. (b) The life histogram shows relative frequency of scale at which the local maxima of the noise image are annihilated as Gaussian blurring proceeds.

the terminating scale of the critical curve of the mode, which is equivalent to the scale of the singular point. We showed that the uniformly distributed data points present a Weibull-like unimodal distribution of the life. The valid cluster can be defined as a cluster with a statistically significant life out of this unimodal distribution. Consequently, the cluster validation is achieved by the statistical rejection method using the unimodal life distribution. The estimated PDF for the uniformly distributed data points is interpreted as an image of the background noise in image analyses.

For the purpose of the statistical selection of the scale for the graph representations in Section 3, we investigate experimentally the distribution of the life of the critical points detected in a two-dimensional noise image with uncorrelated random pixel values as shown in Fig. 4(a). We averaged the frequencies of lives over a hundred of noise images with 128×128 pixels. Figure 4(b) is the averaged histogram of life for local maxima. The obtained life histogram shows an unimodal shape. The critical points with significantly large values of life out of this unimodal distribution can be identified to be valid, because they are distinguishable from the invalid critical points of the noise image.

We can set a critical value of the scale to judge the critical points valid or invalid. Although the computation of such a critical scale requires the parametric model of the life distribution in the strict sense of statistics, the critical scale can be roughly evaluated by the peak and decaying form of life histogram. The peak is found at a small scale relative to the outlying lives if the image contains valid critical points. According to our experimental result, a critical point with a life which is more than a hundred times greater than the peak can be considered to be valid with the statistical confidence level $\alpha > 98.9\%$ under the assumption of uncorrelated random pixel values of a two-dimensional image.

5 Concluding Remarks

The spatial gradient of the Gaussian scale-space image clarifies the hierarchical and topological structure of the image. We presented a scale-selection criterion for the scale-space tree and the scale-space skeleton. These graph representations consist of the critical points of which validity is statistically guaranteed above the selected scale. Our image analysis is wholly integrated in the Gaussian scale-space framework. The scale-space theory has considerable potential for combining multiscale feature detection and statistical validation for image representations.

References

1. Witkin, A.P.: Scale space filtering. In: Proc. of 8th IJCAI, pp. 1019–1022 (1986)
2. Koenderink, J.J.: The structure of images. Biological Cybernetics 50, 363–370 (1984)
3. Lindeberg, T.: Scale-Space Theory in Computer Vision. Kluwer, Boston (1994)
4. Weickert, J., Ishikawa, S., Imiya, A.: Linear Scale-Space has First been Proposed in Japan. Journal of Mathematical Imaging and Vision 10, 237–252 (1999)
5. Zhao, N.-Y., Iijima, T.: A theory of feature extraction by the tree of stable viewpoints. IEICE Japan, Trans. D. J68-D, 1125–1135 (1985) (in Japanese)
6. Lifshitz, L.M., Pizer, S.M.: A multiresolution hierarchical approach to image segmentation based on intensity extrema. IEEE Trans. Pattern Analysis and Machine Intelligence 12(6), 529–540 (1990)
7. Griffin, L.D., Colchester, A.: Superficial and deep structure in linear diffusion scale space: Isophotes, critical points and separatrices. Image and Vision Computing 13(7), 543–557 (1995)
8. Olsen, O.F., Nielsen, M.: Generic events for the gradient squared with application to multi-scale segmentation. In: ter Haar Romeny, B.M., Florack, L.M.J., Viergever, M.A. (eds.) Scale-Space 1997. LNCS, vol. 1252, pp. 101–112. Springer, Heidelberg (1997)
9. Florack, L.M.J., Kuijper, A.: The topological structure of scale-space images. Journal of Mathematical Imaging and Vision 12(1), 65–79 (2000)
10. Kuijper, A., Florack, L.M.J., Viergever, M.A.: Scale space hierarchy. Journal of Mathematical Imaging and Vision 18(2), 169–189 (2003)
11. Sakai, T., Imiya, A.: Gradient structure of image in scale space. Journal of Mathematical Imaging and Vision 28(3), 243–257 (2007)
12. Roberts, S.J.: Parametric and non-parametric unsupervised cluster analysis. Pattern Recognition 30(2), 261–272 (1997)
13. Nakamura, E., Kehtarnavaz, N.: Determining number of clusters and prototype locations via multi-scale clustering. Pattern Recognition Letters 19(14), 1265–1283 (1998)
14. Leung, Y., Zhang, J.-S., Xu, Z.-B.: Clustering by scale-space filtering. IEEE Trans. on Pattern Analysis and Machine Intelligence 22(12), 1396–1410 (2000)
15. Sakai, T., Imiya, A., Komazaki, T., Hama, S.: Critical scale for unsupervised cluster discovery. In: Perner, P. (ed.) MLDM 2007. LNCS (LNAI), vol. 4571, pp. 218–232. Springer, Heidelberg (2007)
16. Muybridge, E.: Animal Locomotion (1887)

Hierarchical Matching Using Combinatorial Pyramid Framework

Luc Brun and Jean-Hugues Pruvot

Université de Caen Basse-Normandie, GREYC CNRS UMR 6072, Image Team
6, Boulevard Maréchal Juin – 14050 Caen Cedex France
{luc.brun,jhpruvot}@greyc.ensicaen.fr

Abstract. A string matching approach is proposed to find a region correspondance between two images. Regions and their spatial relationships are represented by two combinatorial pyramids encoding two segmentation hierarchies. Our matching algorithm is decomposed in two steps: We first require that the features of the two matched regions be similar. This threshold on the similarity of the regions to be matched is used as a pruning step. We secondly require that at least one cut may be determined in each hierarchy such that the cyclic sequence of neighbors of the two matched regions have similar features. This distance is based on a circular string matching algorithm which uses both the orientability of the plane and the hierarchical encoding of the two regions to reduce the computational cost of the matching and enforce its robustness.

1 Introduction

Image correspondence plays a major role in many applications like image indexation, tracking or 3D reconstruction. The correspondence between both images is achieved by matching various primitives of dimension 0 (point), 1 (edge) or 2 (region). The detection of point and edge [1] primitives depends on local information and is thus sensible to noise and missing data. Moreover, the features attached to these primitives only provide local information. On the other hand, segmentation algorithms provide a partition of the image into regions, each region being associated to an area of the image on which complex photometric, geometrical and topological features may be computed.

Feature based region matching algorithms associate to each region a vector based on its photometric or geometric [2] features. Such algorithms neglect the neighbourhood of the two regions which encodes the context of the two regions being matched. These neighbourhoods may be taken into account by using inexact graph matching algorithms [3][4] on the graphs encoding the adjacency relationships between regions in both partitions. However, the application of these algorithms have been limited by the following two reasons:

1. An image, considered as a part of a plane, is an oriented surface. Moreover, the adjacency relationships between the regions of the partition define a planar graph. Most of graph matching algorithms neglect these two properties of a partition which constraint the matching and reduce its time complexity.

2. Any object within an image may be described at different scales. A matching algorithm may thus have to match one region of one image with several regions of the other image, both sets of regions describing a same object.

Wang [3] takes into account the planar structure of a partition by formulating the region correspondence into an inexact matching of two attributed planar graphs. He uses some properties of the planar graph and the topological relationships between regions to find a maximal inexact subgraph matching. However, this method doesn't take advantage of the plane orientation.

The orientation is used by Neuhauss [5] and Lladós [4] which cast the problem of region correspondence between two images into a problem of circular string matching. The two strings being matched encode the boundaries of the matched regions. This explicit use of the orientation drastically reduces the combinatorial complexity of the match. However, both algorithms are sensible to an over segmentation of one of the two images which may split one of the region being matched or one of its neighbor into several sub regions.

This last drawback is addressed by the method proposed by Glantz [6]. This method builds an irregular pyramid from each image and creates one graph from the higher levels of each hierarchy. Each graph is built by connecting the planar graphs encoding the higher levels of the hierarchies by vertical edges. These edges encode the topological relationships between each vertex of one pyramid and its child in the level bellow. The use of hierarchical segmentation algorithms reduces the influence of the over/under segmentation problem and thus increases the number and the quality of the matches. However, the vertical links between the planar graphs destroy the planar structure of the two graphs being matched and thus forbids any use of the good properties of planar graphs. Moreover, the graphs deduced from each pyramid are defined from the connection of several planar graphs and are thus usually large.

The basic idea of our method is to combine the approaches of Glantz and Neuhauss or Lladós. Following the method of Glantz, we first build two irregular pyramids on the two images to be matched. However, instead of building a large graph from each pyramid, we initiate a match by searching in the two hierarchies the couple of regions whose features are similar up to a threshold. Following the method of Neuhauss and Lladós, the boundaries of two candidate regions are matched using a circular string matching algorithm. We take into account both hierarchies by considering the different rewritings of the boundary of each region in its hierarchy. The resulting algorithm is thus a circular string matching algorithm with rewriting rules. The two regions whose features are similar up to the threshold and whose boundary's distance is minimal are considered as good candidates for a larger match.

This paper describes the first step of this matching process which consists to select the two candidate regions for a global match. We first present the combinatorial pyramid framework used to encode both hierarchies in Section 2. The selection of the candidate regions based on their features and our circular string matching algorithm are presented in Section 3. Finally, experiments and results are discussed in Section 4.

2 The Combinatorial Pyramid Framework

A 2D combinatorial map may be understood as a particular encoding of a planar graph where each edge is split into two half-edges called darts. Since each edge of a graph connects two vertices, each dart belongs to only one edge and one vertex. A 2D combinatorial map is formally defined by the triplet $G = (\mathcal{D}, \sigma, \alpha)$ where \mathcal{D} represents the set of darts and σ is a permutation on \mathcal{D} whose cycles correspond to the sequence of darts encountered when turning counter-clockwise around each vertex. Finally α is an involution on \mathcal{D} which maps each of the two darts of one edge to the other one. Given a combinatorial map $G = (\mathcal{D}, \sigma, \alpha)$, its dual is defined by $\overline{G} = (\mathcal{D}, \varphi, \alpha)$ with $\varphi = \sigma \circ \alpha$. The cycles of permutation φ encode the faces of the combinatorial map. In what follows, the cycles of α , σ and φ containing a dart d will be respectively denoted by $\alpha^*(d)$, $\sigma^*(d)$ and $\varphi^*(d)$.

If a combinatorial map $G = (\mathcal{D}, \sigma, \alpha)$ is used to encode a 2D partition, each boundary between two regions is encoded by an edge. Such an edge (say $\alpha^*(d) = (d, \alpha(d))$) encodes an adjacency relationship between $\sigma^*(d)$ and $\sigma^*(\alpha(d))$. The two darts d and $\alpha(d)$ encode the two opposite orientations along the boundary of $\alpha^*(d)$. A σ -cycle $\sigma^*(d) = (d_1, \dots, d_n)$ may thus be understood in two different ways:

- Firstly, since each dart encodes an adjacency relationship, the sequence $(d_i)_{i \in \{1, \dots, n\}}$ determines the cyclic sequence $(\sigma^*(\alpha(d_1)), \dots, \sigma^*(\alpha(d_n)))$ of vertices encountered when turning counter-clockwise around $\sigma^*(d)$. Within a segmentation scheme, this sequence of vertices corresponds to the sequence of regions encountered when turning counter-clockwise around the central region defined by $\sigma^*(d)$.
- Secondly, if we consider that each dart encodes an oriented boundary between two regions, the sequence $\sigma^*(d) = (d_1, \dots, d_n)$ may be considered as the concatenation of the oriented boundaries associated to the darts $(d_i)_{i \in \{1, \dots, n\}}$. This concatenation encodes a counter-clockwise traversal of the boundary of the region encoded by $\sigma^*(d)$.

For example, the two darts of the edge $\alpha_2^*(5) = (5, -9)$ in Fig. □ (c) encode two opposite orientations of the boundary between the regions R_1 and R_2 respectively encoded by the vertices $\sigma_2^*(5) = (5, 16)$ and $\sigma_2^*(-9) = (-9, 20)$. We may associate to the vertex $\sigma_2^*(5)$ the sequence $(\sigma_2^*(9), \sigma_2^*(-21))$ encoding the sequence of regions encountered when turning counter-clockwise around R_1 . Note that in this example, $\sigma_2^*(-21)$ encodes the background of the image.

A combinatorial pyramid is defined by an initial combinatorial map successively reduced by a sequence of contraction or removal operations. Contraction operations are encoded by contraction kernels. A contraction kernel is defined as a forest of the current combinatorial map, each tree of this forest encodes the contraction of a connected set of vertices into a single vertex of the reduced combinatorial map. Contraction kernels may create redundant edges such as empty-self loops and double edges. These redundant edges are removed by

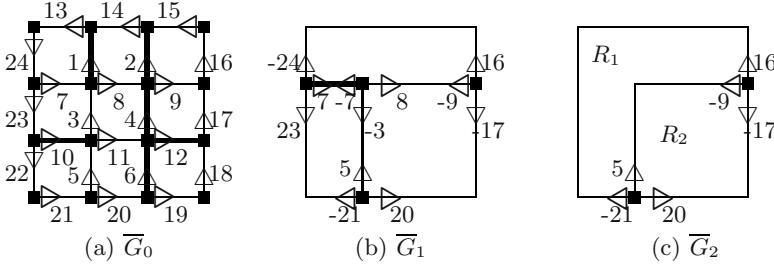


Fig. 1. The first three dual combinatorial map $\overline{G}_0, \overline{G}_1, \overline{G}_2$ of a combinatorial pyramid. The initial map G_0 encodes a 3×3 grid of pixels. Contracted darts are represented by bold lines in (a) and (b).

using respectively empty self loop and double edge removal kernels [7]. These removal kernels are applied successively and are defined as a forest of the dual current combinatorial map. Further details about the construction scheme of a combinatorial pyramid may be found in [7].

A combinatorial pyramid may thus be encoded by a sequence of successively reduced combinatorial maps $P = (G_0, \dots, G_n)$ where G_0 encodes the 4 connected grid of pixels or any other initial partition. Let us consider a level $G_i = (\mathcal{D}_i, \sigma_i, \alpha_i)$ of the pyramid P and a dart d belonging to \mathcal{D}_i . The receptive field of the dart d in G_0 is called the connecting dart sequence of d . Intuitively, this sequence encodes the darts of G_0 combined between levels 0 and i to form the dart d at level i . Since d encodes an oriented boundary between two regions at level i , the embedding of this boundary within G_0 may be retrieved from the connecting dart sequence of d . This sequence called the sequence of boundary darts of d at level i and denoted $SBD_i(d) = d_1 \dots d_p$ is defined by:

$$d_1 = d, d_{j+1} = \varphi_0^{m_j}(\alpha_0(d_j)) \text{ and } \alpha_0(d_p) = \alpha_i(d). \quad (1)$$

where $\overline{G}_0 = (\mathcal{D}_0, \varphi_0, \alpha_0)$ is the dual of the initial combinatorial map and m_j is the minimal integer q such that $\varphi_0^q(\alpha_0(d_j))$ survives at level i or belongs to a double edge kernel. This last condition is tested in constant time using the implicit encoding of combinatorial pyramids [7].

The sequence of boundary darts of the dart 20 at level 2 in Fig. 1(c), is for example equal to $SBD_2(20) = 20.19.18.17$ (Fig. 1(a)). The whole boundary of the vertex $\sigma_2^*(20) = (20, -9)$, encoding the four bottom-right pixels of the 3×3 grid at level 2 , is defined as $SBD_2(20)SBD_2(-9) = 20.19.18.17.-9.-8.-3.-5$.

3 Hierarchical Matching

Let us consider two pyramids $P = (G_0, \dots, G_N)$ and $P' = (G'_0, \dots, G'_{N'})$. Each combinatorial map G_i of P contains a set of vertices. We denote by \mathcal{V}_P the union of all the vertices of P defined between levels 0 and N . In the same way, the set of vertices of P' is denoted $\mathcal{V}_{P'}$. As mentioned in Section 1, our hierarchical matching algorithm aims at finding two vertices in \mathcal{V}_P and $\mathcal{V}_{P'}$ encoding regions

with close features and a minimal distance between their neighbourhoods. We thus want to determine the couple of vertices $(v, v') \in \mathcal{V}_P \times \mathcal{V}_{P'}$ such that:

$$(v, v') = \underset{(w, w') \in \mathcal{C}}{\operatorname{argmin}} \Delta_N(w, w') \text{ with} \quad (2)$$

$$\mathcal{C} = \{(w, w') \in \mathcal{V}_P \times \mathcal{V}_{P'} \mid F(w, w') = \text{true}\} \quad (3)$$

where $\Delta_N(v, v')$ represents the distances between the neighbourhoods of v and v' , and $F(w, w')$ is a boolean condition on w and w' .

The set \mathcal{C} represents our set of candidates for the match. Within our framework Equation 3 is used as a filtering step. The advantages of this filtering step are twofold: Firstly, our criterion F insures that two matched vertices have similar features independently of the distance between their neighbourhoods. Secondly, this filtering reduces the number of couples of vertices on which Equation 2 will be evaluated. Our filtering step and our neighbourhood's distance are respectively described in Section 3.1 and Section 3.2.

3.1 Filtering

Let us associate to each vertex v of \mathcal{V}_P a vector $\text{Feat}(v)$ of M features. Let us suppose that v is defined at level $i \geq 1$ in P and let us consider the set of vertices $RW_i(v) = \{v_1, \dots, v_p\}$ whose contraction in G_{i-1} defines v at level i . The set $RW_i(v)$, called the reduction window of v , is defined as the set of vertices incident to one of the trees of the contraction kernel which builds G_i from G_{i-1} (Section 2). The function $Father$ is defined on \mathcal{V}_P as $Father(w) = v$ for any $w \in RW_i(v)$. By convention any vertex of the top level combinatorial map is its own father. We say that one coordinate j of our vector of feature is increasing iff for any vertex $v \in \mathcal{V}_P$ defined at level $i \geq 1$ in P we have:

$$\forall w \in RW_i(v), \quad \text{Feat}_j(v) \geq \text{Feat}_j(w)$$

Note that any cumulative moment defines an increasing feature along the pyramid.

Let us suppose that the first coordinate $\text{Feat}_1(v)$ of our vector of features is increasing. We first reduce the number of candidates for the match by selecting the set of couples $(v, v') \in \mathcal{V}_P \mathcal{V}_{P'}$ such that:

$$\Delta_{F_1}(v, v') = |1 - \frac{\text{Feat}_1(v')}{\text{Feat}_1(v)}| \leq \epsilon_1 \quad (4)$$

where ϵ_1 is an user defined threshold.

For each vertex $v \in \mathcal{V}_P$ this first filtering step is equivalent to select all vertices v' of $\mathcal{V}_{P'}$ such that $\text{Feat}_1(v') \in [(1 - \epsilon_1)\text{Feat}_1(v), (1 + \epsilon_1)\text{Feat}_1(v)]$. This filtering step may be achieved efficiently using the increasing property of the first feature's coordinate and the hierarchical relationships of P' encoded by the reduction window and father functions.

The first filtering step defined by equation 4 reduces the number of potential candidates for the match based on only one feature. We refine this first filtering

by using the remaining features of each vertex. However, since the different features associated to a vertex vary within different intervals, a normalisation step is required before computing the distance between the features of two vertices. Given two vertices v and v' we define the distance between $Feat(v)$ and $Feat(v')$ as the infinite norm of a vector f defined as:

$$\forall i \in \{1, \dots, M\} \quad f_i(v, v') = 1 - \frac{\min(Feat_i(v), Feat_i(v'))}{\max(Feat_i(v), Feat_i(v'))} \quad (5)$$

Our final boolean criterion encoding the set \mathcal{C} of filtered couple of vertices (equation 3) is thus equal to:

$$F(w, w') = \Delta_{F_1}(v, v') \leq \epsilon_1 \text{ and } \|f\|_\infty \leq \epsilon_1.$$

Both tests being applied sequentially.

Note that if the first feature $Feat_1(v)$ is positive then: $\Delta_{F_1}(v, v') \leq \epsilon_1$ implies that $f_1(v, v') \leq \epsilon_1$. Our first filtering step ($\Delta_{F_1}(v, v') \leq \epsilon_1$) may in this case be interpreted as a restriction of our second test ($\|f\|_\infty \leq \epsilon_1$) to the first coordinate of f . Such an interpretation is loss if we use the Euclidean norm rather than the infinite one. The Euclidean norm may however be preferred if one wants to allow a phenomenon of compensation between the feature's distances f_i .

3.2 Distance between Hierarchical Neighbourhoods

Given a pyramid P , let us consider a vertex $v \in \mathcal{V}_P$ which survives up to level l_v . Let us denote by $\sigma_{l_v}^*(d) = (d_1, \dots, d_q)$ the σ -cycle of G_{l_v} associated to v . The embedding of the boundary of v within G_0 , denoted B_v is defined as the concatenation of the sequence of boundary darts $(SBD_{l_v}(d_j))_{j \in \{1, \dots, q\}}$ (Section 2). Since all the darts of a sequence of boundary darts belong to $G_0 = (\mathcal{D}_0, \sigma_0, \alpha_0)$, any vertex of the hierarchy may be considered as a word built on the alphabet \mathcal{D}_0 . The neighbourhood of v in G_0 is defined as the set of vertices of G_0 adjacent to the eventual over-segmentation of v . This set, denoted $N^0(v)$ may be formally defined by: $N^0(v) = \{\sigma_0^*(\alpha_0(d)), d \in B_v\}$.

Let us consider two pyramids P , P' and two vertices $(v, v') \in \mathcal{V}_P \times \mathcal{V}_{P'}$ with $B_v = d_1, \dots, d_n$ and $B_{v'} = d'_1, \dots, d'_{n'}$. Using the orientation of $N^0(v)$ and $N^0(v')$ in G_0 and G'_0 , the neighbourhood's distance between v and v' at level 0 is defined as the edit distance between B_v and $B_{v'}$. If q denotes the number of symbols matched between B_v and $B_{v'}$, the total cost of the match between v and v' is defined as:

$$\Delta(B_v, B_{v'}) = \sum_{i=0}^q \delta(d_{\phi_1(i)}, d'_{\phi_2(i)}) + (n + n' - 2q)K \quad (6)$$

where $\delta(., .)$ is a distance function between darts and K is the default cost for a removal operation within either B_v or $B_{v'}$. The values of $\phi_1(i)$ and $\phi_2(i)$ correspond to the indexes within B_v and $B_{v'}$ of the two i^{th} darts being matched.

As mentioned in Section 2, each dart of a combinatorial map may be interpreted both as an encoding of the adjacency between two regions and as a boundary between the same two regions. Therefore, if we consider two darts $d \in G_i$ and $d' \in G'_j$ of P and P' , the distance $\delta(d, d')$ between these two darts may incorporate features of the vertices $\{\sigma_i^*(d), \sigma_i^*(\alpha_i(d)), \sigma_j'^*(d'), \sigma_j'^*(\alpha_j(d'))\}$ and geometrical features of the boundaries associated to the darts d and d' at levels i and j . Within our framework, since the distance between the features of v and v' has been tested during the filtering step we define $\delta(d, d')$ from the features of the vertices $\sigma_i^*(\alpha_i(d))$, $\sigma_j'^*(\alpha_j(d'))$ and the features of the two boundaries encoded by d and d' at level i and j (Section 4).

All the darts of the sequences B_v and $B_{v'}$ are defined in G_0 . Therefore, the distances $\delta(.,.)$ in Equation 6 are evaluated at level 0 and the whole comparison of the neighbourhoods of v and v' is made in G_0 and G'_0 . As mentioned in Section 11, the neighbourhoods of v and v' may not correspond within the two initial partitions encoded by G_0 and G'_0 . However, given the two pyramids P and P' respectively built on G_0 and G'_0 , we may consider the set of cuts of P (resp. P') which pass below v (resp. v'), i.e. such that v (resp. v') is either a single region or is over-segmented within the cut. The basic assumption of this paper is that if v and v' correspond to a same “ideal” region there should be at least one such cut for each pyramid in which the neighbourhoods of v and v' are similar.

Let us consider a sub sequence $d_l \dots d_m$ of B_v such that the set of vertices $\{\sigma_0^*(\alpha_0(d_l)), \dots, \sigma_0^*(\alpha_0(d_m))\}$ is merged into a single vertex at a level $l_1 \leq l_v$. All the darts encoding the adjacency relationships between $\{\sigma_0^*(\alpha_0(d_l)), \dots, \sigma_0^*(\alpha_0(d_m))\}$ should thus be either contracted or removed as empty self loop at level l_1 . Moreover, the set of vertices $\{\sigma_0^*(d_l), \dots, \sigma_0^*(d_m)\}$ defines an over-segmentation of v and should be merged into a single vertex at a level $l_2 \leq l_v$. If we consider the sequence $d_l \dots d_m$ at the level $L = \max(l_1, l_2)$, the darts $d_l \dots d_m$ encode a contiguous sequence of boundaries between two regions and may thus be interpreted as double edges [8]. These darts should thus belong to the sequence of boundary dart $SBD_L(d_k)$ of a dart $d_k \in B_v$. Conversely, a sequence of boundary darts $SBD_L(d_k) = d_l \dots d_m$ with $L \leq l_v$ and $d_k \in B_v$ encodes a sequence of double edge darts which may be produced only by the merge of the vertices $\{\sigma_0^*(d_l), \dots, \sigma_0^*(d_m)\}$ and $\{\sigma_0^*(\alpha_0(d_l)), \dots, \sigma_0^*(\alpha_0(d_m))\}$. The sequence of boundary darts allows thus to retrieve all the merge operations between the vertices of $N^0(v)$ defined before the level l_v . Since two cuts producing the same merge operations on $N^0(v)$ produce the same local configuration in the neighbourhood of v , the sequence of boundary darts $SBD_i(d)$ with $i \leq l_v$ and $d \in B_v$ allows us to retrieve all the neighbourhoods of v produced by the different cuts of P .

For any sequence of boundary dart $SBD_i(d_j) = d_j \dots d_k$ with $i \leq l_v$ and $d_j \in B_v$ let us consider the rewriting rule $\psi_{j,k}^i : d_j \dots d_k \mapsto d_j$ which replaces in B_v , the sequence of darts $d_j \dots d_k$ defined in G_0 by the single dart d_j considered as a dart of G_i . In order to avoid confusions between the dart d_j considered in G_0 and in G_i , we use the same notation for the name of the rule $\psi_{j,k}^i$ and its result d_j

in G_i . We have thus: $\psi_{j,k}^i : d_j \dots d_k \mapsto \psi_{j,k}^i$. From a geometrical point of view, the oriented boundary encoded by the dart $\psi_{j,k}^i$ is defined as the concatenation of the oriented boundaries encoded by $d_j \dots d_k$ at level 0. Let us consider the set Σ of such rules defined on B_v and $\Sigma(B_v)$ the set of rewritings of B_v . In the same way, let us consider the set of rewritings of $B_{v'}$ denoted as $\Sigma'(B_{v'})$. Both $\Sigma(B_v)$ and $\Sigma'(B_{v'})$ encode the neighbourhoods of v and v' among the different cuts of P and P' . The neighbourhood's distance between v and v' is thus defined as:

$$\Delta_N(v, v') = \min_{(m, m') \in \text{Rot}(\Sigma(B_v)) \times \Sigma'(B_{v'})} \Delta(m, m')$$

where Δ is defined according to Equation 6 and $\text{Rot}(\Sigma(B_v))$ represents all the circular permutations of the strings contained in $\Sigma(B_v)$.

Given a rotation, such a distance may be computed efficiently [9] using the following recursive equation:

$$\Delta(0, 0) = 0, \text{ and } \Delta(i, j) = \min \left(\begin{array}{l} \Delta(i-1, j-1) + \delta(d_i, d_j), \\ \min_{\psi_{k,i}^l \in \Sigma, \psi'^{l'}_{k',j} \in \Sigma'} \Delta(k-1, k'-1) + \delta(\psi_{k,i}^l, \psi'^{l'}_{k',j}), \\ \Delta(i-1, j) + K, \\ \Delta(i, j-1) + K \end{array} \right)$$

where $\psi_{k,i}^l$ and $\psi'^{l'}_{k',j}$ denote two rewriting rules of Σ and Σ' . The indexes k and k' are respectively lower than i and j and the levels l and l' must be respectively lower than l_v and $l_{v'}$. Note that the distance $\delta(d_i, d_j)$ in the above equation is evaluated at level 0.

4 Experiments

Given a pyramid P and a vertex $v \in \mathcal{V}_P$, let us denote by R_v the geometrical region associated to v and by ∂R_v the boundary of R_v . The vector of features $\text{Feat}(v)$ (Section 3.1) used in our experiments is composed of $M = 6$ features.

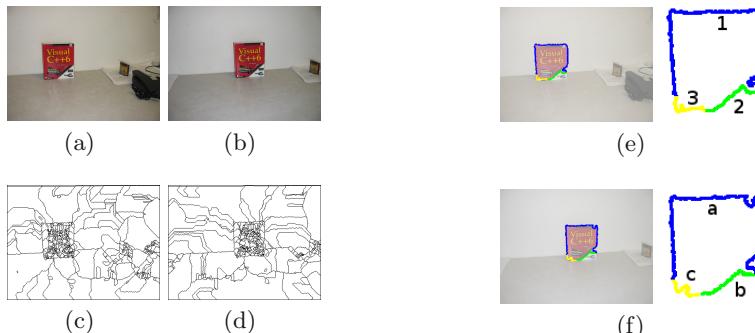


Fig. 2. Two matched boundaries (e) and (f) determined from two pyramids obtained from the initial partitions (c) and (d) and the images (a) and (b).

$Feat_1(v)$ encodes the size of R_v while $Feat_2(v)$ to $Feat_4(v)$ encode the three colour channels of the mean colour of R_v . The two last features $Feat_5(v)$ and $Feat_6(v)$ respectively encode the size of ∂R_v and the mean value of the colour gradient computed along it.

Given a dart d defined at level i , we characterise the shape of the oriented boundary associated to d by a function which maps each discrete point of the boundary to the curvature [10] of the boundary at this point. We obtain thus a vector of curvature points that we compress into 8 features using the Legendre's moments [11]. We add to these 8 geometrical features 3 colour features corresponding to the mean colour of the vertex $\sigma_i^*(\alpha_i(d))$ (Section 3.2). The distance function δ (Section 3.2) between two darts is then defined as the Euclidean norm of a vector f using a normalisation step equivalent to the one used in Equation 5 but applied on the features of the two darts.

One result of our algorithm on a couple of real images representing a same scene with different view points is presented in Fig. 2. The base level combinatorial maps of the two pyramids $P = (G_0, \dots, G_n)$ and $P' = (G'_0, \dots, G'_{n'})$ encode a watershed of the input images [12]. The construction scheme of the remaining levels of the pyramids is described in [13]. The two matched vertices v and v' encode the upper left part of the book in both images. Using our rewriting rules, the sequences B_v and $B_{v'}$ encoding the shape of the book in the base level combinatorial maps (Fig. 2(c), (d)) have been grouped into only 3 darts (Fig. 2(e), (f)): The darts, separating the book and the wall ($1 \leftrightarrow a$), the white and the red part of the book ($2 \leftrightarrow b$) and the book and the table ($3 \leftrightarrow c$). All these darts are matched at different levels in the pyramids P and P' . For example, the dart 1 belongs to the 838th level of P while a belong to the 961th level of P' .

Fig. 3 illustrates the behavior of our algorithm on a more complex example. The sequences B_v and $B_{v'}$ encoding the shape of the cars in the base level combinatorial maps (Fig. 3(c), (d)) have been grouped into 4 darts (Fig. 3(e), (f)). We used the same convention for the matches than in Figure 2(a,1),(b,2)....

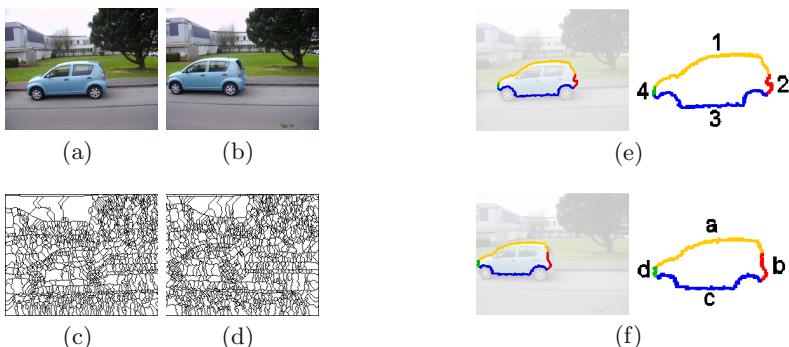


Fig. 3. Two matched boundaries (e) and (f) determined from two pyramids obtained from the initial partitions (c) and (d) and the images (a) and (b).

5 Conclusion

We have presented in this paper a method which matches two similar regions R and R' in two hierarchies P and P' . This method uses the features of R and R' and their oriented neighbourhoods. The distance between the oriented neighbourhoods of R and R' is defined as the minimal distance between the neighbourhoods of R and R' in all the cuts of P and P' which include them. The result of this new notion of similarity is a match between two regions together with a grouping of their boundary's basic elements into larger groups. The resulting boundaries are finally aligned. This last features should allows us to expand the match in order to find the largest common sub-map, within two hierarchies.

References

1. Kumar, S., Sallam, M., Goldgof, D.: Matching point features under small nonrigid motion. *Pattern Recognition* 34, 2353–2365 (2001)
2. Kaygin, S., Bulut, M.M.: Shape recognition using attributed string matching with polygons vertices as the primitives. *Pattern Recognition Letters* 23, 287–294 (2002)
3. Caihua, W., Keiichi, A.: Region correspondence by inexact attributed planar graph matching. In: 5th International Conference on Computer Vision, June 1995, pp. 440–447 (1995)
4. Lladós, J., Martí, E., Villanueva, J.: Symbol recognition by error-tolerant subgraph matching between region adjacency graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 1137–1143 (2001)
5. Neuhaus, M., Bunke, H.: An error-tolerant approximate matching algorithm for attributed planar graphs and its application to fingerprint classification. In: Fred, A., Caelli, L.M., Duin, R.P.W., Campilho, A.C., de Ridder, D. (eds.) *SSPR&SPR 2004*. LNCS, vol. 3138, pp. 180–189. Springer, Heidelberg (2004)
6. Glantz, R., Pelillo, M., Kropatsch, W.G.: Hierarchical matching of panoramic images. In: *ICIP 2003*, p. 328. IEEE Computer Society, Los Alamitos (2003)
7. Brun, L., Kropatsch, W.: Construction of combinatorial pyramids. In: Hancock, E., Vento, M. (eds.) *GbRPR 2003*. LNCS, vol. 2726, pp. 1–12. Springer, Heidelberg (2003)
8. Brun, L.: Traitement d'images couleur et pyramides combinatoires. Habilitation à diriger des recherches, Université de Reims (2002)
9. Gdalyahu, Y., Weinshall, D.: Flexible syntactic matching of curves and its application to automatic hierarchical classification of silhouettes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21, 1312–1328 (1999)
10. Malgouyres, R., Brunet, F., Fourey, S.: Binomial convolutions and derivatives estimation from noisy discretizations. In: Coeurjolly, D. (ed.) *Proceedings of DGCI 2008*. LNCS, vol. 4992, Springer, Heidelberg (2008)
11. Shen, J., Shen, D.: Orthogonal legendre moments and their calculation. In: *ICPR 1996*, vol. 2, p. 241 (1996)
12. Brun, L., Mokhtari, M., Meyer, F.: Hierarchical watersheds within the combinatorial pyramid framework. In: Andrès, É., Damiand, G., Lienhardt, P. (eds.) *DGCI 2005*. LNCS, vol. 3429, pp. 34–44. Springer, Heidelberg (2005)
13. Pruvot, J.H., Brun, L.: Scale set representation for image segmentation. In: Escalano, F., Vento, M. (eds.) *GbRPR*. LNCS, vol. 4538, pp. 126–137. Springer, Heidelberg (2007)

Multi-cue Facial Feature Detection and Tracking

Jingying Chen and Bernard Tiddeman

School of Computer Science, University of St Andrews
Fife, Scotland, UK KY16 9SX
{jchen,bpt}@cs.st-and.ac.uk

Abstract. An efficient and robust facial feature detection and tracking system is presented in this paper. The system is capable of locating a human face automatically. Six facial feature points (pupils, nostrils and mouth corners) are detected and tracked using multiple cues including facial feature intensity and its probability distribution, geometric characteristics and motion information. In addition, in order to improve the robustness of the tracking system, a simple facial feature model is employed to estimate the relative face poses. This system has the advantage of automatically detecting the facial features and recovering the features lost during the tracking process. Encouraging results have been obtained using the proposed system.

1 Introduction

Facial feature detection and tracking is important in vision related applications such as human machine interaction [1], facial expression analysis [2], facial image transformation [3] and head pose tracking [4]. These applications need to track the facial features robustly and efficiently. A robust facial feature tracking system should incorporate automatic feature detection, tracking failure detection and feature recovery capability. However, the high variability of face morphology, head motion and complex background together with unknown and variable illumination conditions make the detection and tracking task difficult and complex. Hence, we propose a system capable of automatically locating a human face, analyzing its orientation, and then detecting and tracking six facial feature points, (pupils, nostrils and mouth corners), in a real time video. Integrating detection and tracking into a single system is important for recovering features after tracking failure e.g. due to temporary occlusion of the tracked features.

The facial feature detection and tracking literature includes image-based approaches [5, 6], template-based approaches [7, 8, 9], appearance-based approaches [10, 11] and motion-based approaches [12]. Each of these approaches has its own strengths and limitations. Image-based approaches use color information, properties of facial features and their geometric relationships to locate facial features. Yang and Stiefelhagen [5] presented a technique for tracking based on human skin color. This approach is in general very fast, however, color alone does not provide enough reliable information to track facial features. Stiefelhagen et al. [6] used

color information and certain geometric constraints on the face to detect and track six facial feature points (pupils, nostrils and mouth corners) in real time for lip reading. This method works properly under good lighting conditions, however the mouth corners may drift away when the illumination changes. Template based approaches are usually applied to intensity images where a predefined template of facial feature is matched against image blocks. Tian et al. [7] used multiple state templates to track the facial features. Feature point tracking together with masked edge filtering is used to track the upper facial features. The system requires that templates be manually initialized in the first frame of the sequence, which prevents it from being automatic. Kapoor and Picard [8] used eyebrow and eye templates to locate upper facial features in a real time system. However, specialized hardware (an infrared sensitive camera equipped with infrared LEDs) is needed to produce the red eye effect in order to track the pupils. Matsumoto and Zelinsky [9] detected the facial features using an eye and mouth template matching method, which was implemented using the IP5000 image processing board.

Appearance-based approaches use facial models derived from a large amount of training data. These methods are designed to accommodate possible variations of human faces under difference conditions. Cootes et al. [10] proposed active appearance models (AAM) and Matthews and Baker [11] improved the performance of the original AAM. However, these methods need large amounts of delineated training data and involve relatively expensive computations. Also, the AAM fitting requires expensive computations which make the real time tracking difficult. Cristinacce and Cootes [12] proposed Constrained Local Model (CLM) for feature detection and tracking, they used a joint shape and texture appearance model to generate a set of region template detectors. The model is fitted to an unseen image in an iterative manner by generating templates using the joint model and the current parameter estimates, correlating the templates with the target image to generate response images and optimising the shape parameters so as to maximise the sum of response. In their method, Viola and Jones's [13] features are used to detect face. Within the detected face region they applied smaller Viola and Jones's feature detectors constrained using the Pictorial Structure Matching (PSM) method [14], to detect initial feature points. The PSM combines feature responses and shape constraints, which is very efficient due to the use of pairwise constraints and a tree structure. They claimed their proposed method is more robust and accurate than the original AAM. The method described here is similar to this method in terms of detection quality, but requires less processing time per frame. Bourel et al. [15] proposed a motion based facial feature point tracking system. In their method a Kanade-Lucas-Tomasi (KLT) tracker is employed and robust results have been obtained. However, manual initialization is required.

The facial feature detection and tracking approach using a single cue about the image sequence is insufficient for reliable performance. A robust tracking system should use as much knowledge about the image sequence as possible to handle all sources of variability in the environment. Hence, we propose to use the multi-cue of Haar-like features, intensity and its probability distribution, geometry constraints, motion and a

facial feature model to build a robust facial feature tracking system. In this paper, the proposed approach locates a human face without any artificial marks on it. This system is capable of detecting and tracking six facial features (i.e. two pupil centers, two nostril centers and two mouth corners) automatically when a human face appears in front of the camera. A simple facial feature model (locations of the feature points) is used to detect tracking failure and recover from it.

The outline of the paper is as follows. The proposed facial feature detection and tracking are presented in Section 2 and Section 3, respectively. Section 4 describes the experimental results while Section 5 presents the conclusions.

2 Facial Feature Detection

In the proposed approach, a face is first detected, which relies on a boosting algorithm and a set of Haar-like features. Then the pupils are searched for inside the face area based on their intensity characteristics and Haar-like features. Next, the mouth is found using the pupil positions and mouth intensity probability distribution. Finally the nostrils are located using their intensity and geometric constraints. The detail of the detection procedure is given below.

2.1 Face Detection

Viola and Jones's [13] face detection algorithm, based on Haar-like features is used to detect a face. Haar-like features encode the existence of oriented contrast between regions in the image. A set of these features can be used to encode the contrast exhibited by a human face and their special relationships. In Viola and Jones's method, a classifier (i.e. a cascade of boosted classifier working with Haar-like features) is trained with a few hundreds of sample views of face and non-face examples, they are scaled to the same size, i.e. 24x24. After the classifier is trained, it can be applied to a region of interest in an input image. To search for the face, one can move the search window across the image and check every location using the classifier.

After the face is detected, Principal Component Analysis (PCA) [16] is employed within face region for estimation of the face direction. In 2D shape, PCA can be used to detect principal directions of the spatial shape. Since faces are approximately symmetric and there are many edge features around eyes, the first principal axis indicates the upright direction of the face while the second principal axis gives the direction of eyes (i.e. the line connecting the two pupils). Figure 1 shows the face edge map and the principal directions of the edge map.



Fig. 1. Principal axes of face skin region. The major axis Vy represents face direction while the minor axis Vx represents eye direction.

After obtaining the eye direction, one can rotate the face until the eye direction is parallel to the horizontal axis using the middle point of the line connecting the two pupils as the rotation centre, which facilitates the following pupil, nostril and mouth corner detection.

2.2 Eye Detection

2.2.1 Eye Region Localization

After the face is detected and aligned, the eye region can be located based on the face vertical edge map. Pixel intensities change in the vertical direction more near the eyes than in the other parts of the face (e.g. the eyebrows and the side of face boundary). Hence it is reasonable to use a vertical edge map to decrease false positive eye detection. First, a horizontal integral projection [17] is used on the upper face vertical edge map to estimate the vertical position of the eyes. The vertical position of the eyes is usually the global maximum of the horizontal integral projection in the upper face. According to the vertical position and the height of the face, we can estimate the vertical eye region (See Figure 2).

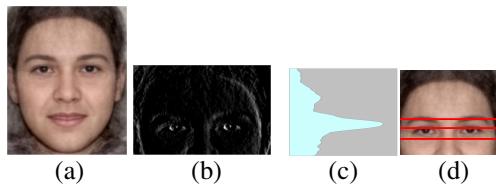


Fig. 2. Vertical eye region detection, (a) face image, (b) upper face vertical edge map, (c) horizontal projection performed on image(b), and (d) estimated vertical eye region

Second, a vertical projection is performed on the face vertical edge map to estimate the right and left boundary of face which correspond the two peaks of projection values (See Figure 3).

Finally, the eye region can be located based on the vertical eye region and right and left face boundary [18] (see Figure 4).

2.2.2 Pupil Detection

In order to improve the accuracy and efficiency of detection, eyes are searched for within the obtained eye region instead of the entire face, which decreases false positive and speeds up the detection process. Similar to face detection described above, eyes are found using a cascade of boosted tree classifiers with Haar-like features. A statistical model of the eyes is trained in this work. The model is made of a cascade of boosted tree classifiers. The cascade is trained on 1000 eye and 3000 non-eye samples of size 18x12. The training set contains different facial expression and head rotation. The 18x12 window moves across the eye region and each sub-region is classified as eye or non-eye. An example of the eye detection is shown in Figure 5.

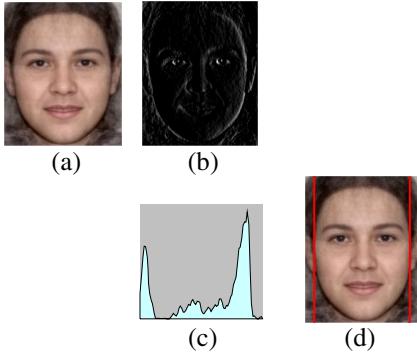


Fig. 3. Face boundary detection, (a) face image, (b) face vertical edge map, (c) vertical projection performed on image(b), and (d) estimated right and left face boundary



Fig. 4. Estimated eye region

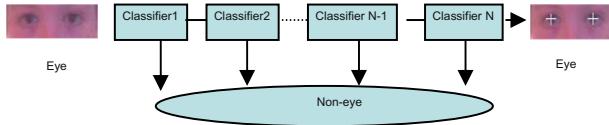


Fig. 5. Example of the eye detection using a cascade of boosted tree classifiers

2.3 Mouth Corners Detection

An estimated mouth region can be obtained using the pupil positions and face size. Then a novel approach based on entropy analysis of the partial histogram within the mouth region is proposed to segment the mouth. In this approach, the entropy E_j is iteratively calculated according to different parts of the histogram, until its value is greater than a threshold E_{th} which is found from the training data.

$$E_j = \sum_{i=0}^j H(i) \log H(i) \quad j=0, \dots, n. \quad n \in (1, 255)$$

where i and $H(i)$ are histogram index and value respectively. When $E_j > E_{th}$, j is used as a threshold to segment mouth. Mouths generally have lower intensities than neighbouring pixels and contain a relatively fixed proportion of the information within the mouth region, hence it is reasonable to segment them based on the threshold chosen from partial histogram entropy, which is insensitive to illumination variations. For example, in Figure 6 one can see that the two mouth images under different illumination conditions can be segmented correctly because the segmentation does not depend on the absolute intensity. Finally, the mouth corner positions are estimated based on the largest connected region of the segmented image (see Figure 6(b)). Extremities of the bright areas around the left and right parts are searched for as mouth corners.

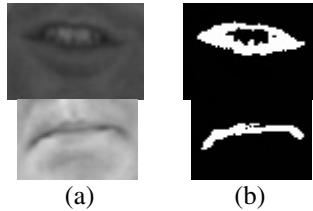


Fig. 6. Mouth corner detection in the segmented image, (a) mouth intensity and (b) segmented images

2.4 Nostrils Detection

The nostrils appear dark relative to the surrounding area of the face under a wide range of lighting conditions. As long as the nostrils are visible, they can be found by searching for two dark regions, which satisfy certain geometric constraints. Here, the search region is restricted to an area below the pupils and above the mouth. Then the automatic thresholding, with a threshold corresponding to the lower 5 percent of the local histogram for the search window is applied. Then, the centers of the dark regions are computed as the centers of nostrils (see Figure 7).



Fig. 7. The illustration of nostrils detection

3 Facial Feature Tracking

Once the feature points have been detected, the Lucas-Kanade (LK) algorithm [19] is performed to track the pupils and nostrils. The algorithm detects the motion through the utilization of optical flow. Since the mouth has relatively higher variability (i.e. closed mouth or open mouth, with/without teeth appearance) compared to the eyes and nose, mouth corners are tracked based on the segmented mouth image. Extremities of the bright areas are searched for around the previously found left and right corners.

In order to build a robust tracking system, the system has to be able to detect tracking failure and recover from it. In this paper, the face is assumed to be planar and a simple facial feature model is used comprising five facial feature points: two centres of pupils, two nostrils and the centre of the mouth. The mouth centre is more reliable and less deformable than the mouth corners.

For each pixel (x_1, y_1) in a given image frame I_1 and the corresponding image point (x_2, y_2) in another image frame I_2 , the relative orientation of two face poses is estimated using the basic projection equation of a weak perspective camera model for planar 3D object points[20].

$$\begin{pmatrix} x_2 - x_{c2} \\ y_2 - y_{c2} \end{pmatrix} = M_2 \times M_1^{-1} \begin{pmatrix} x_1 - x_{c1} \\ y_1 - y_{c1} \end{pmatrix}$$

where M_1 and M_2 are the projection matrices for image I_1 and I_2 respectively, and (x_{c1}, y_{c1}) and (x_{c2}, y_{c2}) are the projection points of the same reference point (X_c, Y_c, Z_c) in the image I_1 and image I_2 . This equation is the fundamental weak perspective homographic projection equation that relates image projections of the same 3D points in two images with different face poses. The homographic matrix $P = M_2 \times M_1^{-1}$ gives the relative orientation between the two face poses [21]. Instead of using all the feature points and some sort of least-square pose fitting method [22], a minimal subset of feature points is used to estimate the pose. So long as one subset of good, accurate measurements exists, the rest of the feature points can be ignored and gross errors will have no effect on the tracking performance. The selection of a good subset can be done within the RANSAC regression paradigm [23]. The computed pose helps the tracking system to detect tracking failure and improves the robustness of the tracking system.

3.1 Tracking Failure Detection and Recovery

In this system, the failure detection includes two steps. First, all the found feature points are checked to see if they lie within the face region and satisfy certain constraints inherent in facial geometry. If not, the model points are projected back onto the image using the computed pose. In the case of mild occlusion, the lost feature points can be recovered (see Figure 9). Second, if the average distance between the back-projected model points and the actual found points exceeds a certain threshold, it is considered a failure.

Once the tracking failure has been detected, the feature points have to be searched for again. The failure recovery can be solved using the previously found pose just before the failure occurs.

If a pupil is lost during the tracking process, a search window is computed. Its center and size are chosen based on the previously found pose. The search window center is the previous position of the pupil and its size is proportional to the Euclidean distance between the two last known pupil positions. This can scale the search window automatically when the person gets closer or further from the camera. Then, the pupil detection described in the previous section is applied within the search window. The nostril recovery is based on the automatic detection of a dark region within a search window, the center of the dark region is computed as the recovered nostril center. The mouth corners are recovered based on entropy analysis of the partial histogram within the mouth region (see Figure 6, as described in the mouth corner detection).

4 Experimental Results

The proposed method has been implemented using C++ under MS windows environment with Pentium M 715 processor and tested with both static face databases and live video sequences.

4.1 Detection Results

Four face databases are used here. The first database includes 181 face images with standardized pose, lighting and camera setup. The second database includes 121 un-standardized mobile phone face images with different lighting, pose and face expression. The third database is an un-standardized set of face images from the internet containing 200 face images including different skin color, various illuminations and face poses and expressions. The final database is publicly available and known as the BIOID database (<http://www.bioid.com>) which consists of 1521 images of front faces under cluttered background and various illuminations. The detection rates (i.e. images with successful detected eyes, nostrils and mouth corners relative to the whole set of the database images) of the eyes, nostrils and mouth corners are 96%, 91% and 92%. Examples of the detected facial features using the proposed method are shown in Figure 8. The white crosses represent the detected features.

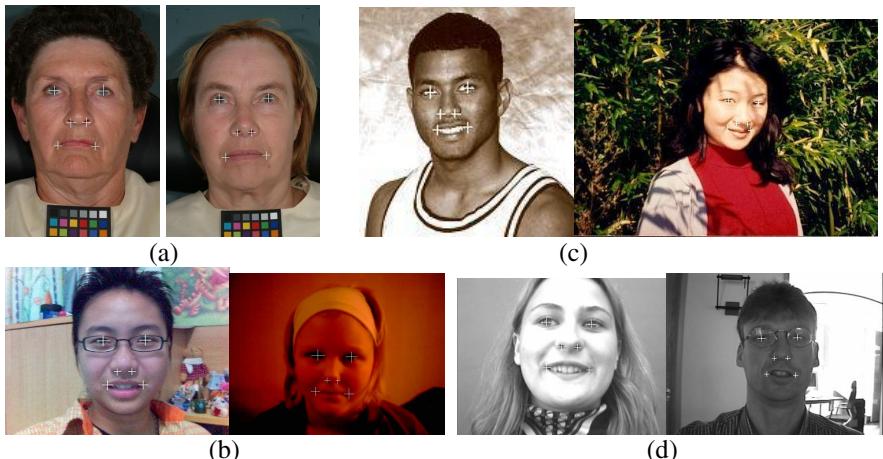


Fig. 8. The results of detected eyes, nostrils and mouth corners from four face image databases. (a) The database of highly standardised photos, (b) the database of un-standardized mobile phone image, (c) un-standardized images from the internet and (d) images from the BIOID database.

From these results, one can see that the proposed approach can detect the eyes and mouth corners accurately under various illumination conditions. However, false detection could exist under some circumstance (see Figure 9). One can notice that the eyebrows and eyes are quite similar and right nostril is invisible from Figure 9 (a) and the moustache occludes the mouth from Figure 9 (b).

To compare the proposed method with other methods, the publicly available BIOID database is used here which was specifically designed to capture faces in realistic authentication conditions and a number of methods have been evaluated on this data set [24, 25, 26]. An eye distance measure (i.e. D_e) introduced by Jersorsky et al. [24] is adopted here, which records the maximum displacement error (the displacement

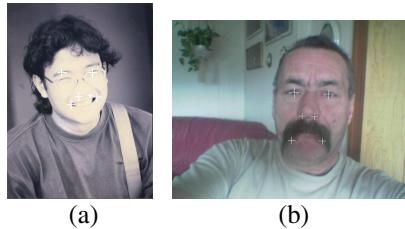


Fig. 9. False detection on eyes and nostrils(a) and mouth(b)

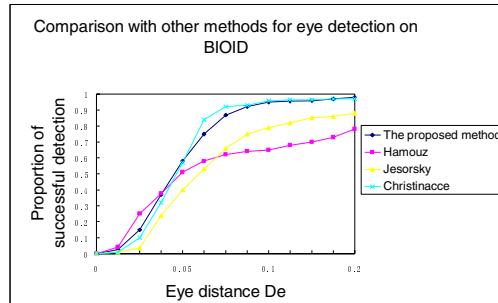


Fig. 10. Comparison with other methods for eye detection on BIOID

error is calculated as the distance between the manually determined positions of the feature points and the detected feature points) between both eyes, normalised by the true eye separation. The comparison results are given in Figure 10. From this figure, one can see that 58% and 94% success rates are obtained when D_e reaches 0.05 and 0.1 using our proposed method. In [24], Jersorsky et al. used a face matching method based on the Hausdorff distance followed by a Multi-Layer Perceptron eye finder. They achieve 40% and 79% success rate with D_e equal to 0.05 and 0.1. Hamouz et al. [25] used a method combining Gabor based feature detection to produce a list of face hypotheses, and which are then tested using a SVM face model. They also presented eye detection results on BIOID, they obtain 51% and 62% success rates when D_e approaches 0.05 and 0.1. Cristinacce et al. [26] presented a Pairwise reinforcement of feature responses approach combined with Active appearance model to detect facial feature points. They acquire 57% and 96% success rates with D_e equal to 0.05 and 0.1 on BIOID. Their method can provide high success rate, however, it needs ~1400ms to search a BIOID image using 500Mhz PII processor. The approach described here requires only ~220ms to search a BIOID image.

4.2 Tracking Results

The tracking experiments have been performed using both live data capture and pre-recorded video sequences. Tracking results from two test sequences are given below

(see Figure 10 and 11). Both sequences were captured using an inexpensive web camera with a resolution of 320 x 240 at 25 frames per second. The white crosses represent the tracked points.

From these results, one can see that the proposed approach can track the six feature points accurately when the person is moving or rotating his head, even in the case of temporary occlusion thanks to the use of the simple model which predicts occluded points during the tracking process. Figure 13 gives an estimate of the tracking accuracy of the proposed tracking system. The measurements were taken using 1027 frames of six subjects in this experiment. Manually corrected positions of the feature points were used as reference for measuring displacement error. The error was computed as the Euclidean distance between the reference points and the points obtained

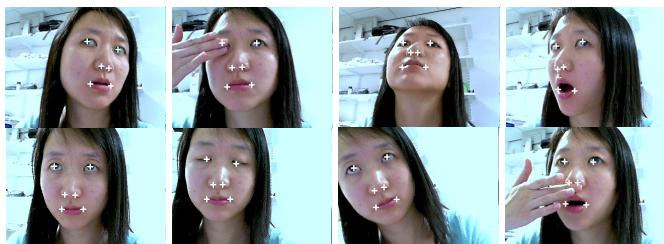


Fig. 11. Tracking results for sequence 1



Fig. 12. Tracking results for sequence 2

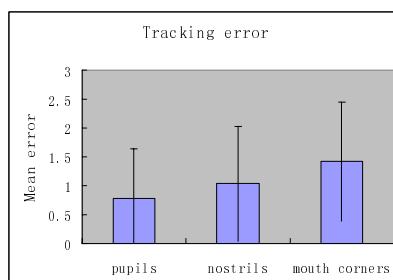


Fig. 13. The average and standard deviation of the distances between the tracked feature points and manually corrected feature points of the pupils, nostrils and mouth corners

by the proposed method. The average and standard deviations of the distances were computed across all pupils, nostrils and mouth corners during the tracking processing (see Figure 13).

The performance of the proposed approach is very good, which can cope with large angle head rotation, different facial expressions and mild occlusion of the face. The approach is fully automatic and Lucas-Kanade is a computationally efficient motion detector. They can be easily implemented in a real time system. On the other hand, a simple facial feature model (locations of the five feature points) is used to improve the system robustness, its simplicity needs inexpensive computation. Hence, the proposed tracking system is efficient and robust, suitable for putting into practice.

5 Conclusions

A multi-cue facial feature detection and tracking system is proposed in this paper, which detects a human face using a boosting algorithm and a set of Haar-like features, determines the face orientation using PCA, locates the pupils based on their intensity characteristics and Haar-like features, finds the mouth corners from the mouth intensity probability distribution, estimates the nostrils based on their intensity and geometric constraints and tracks the detected facial points using optical flow based tracking. The system is able to detect tracking failure using constraints derived from a facial feature model and recovery from it by searching for one or more features using the feature detection algorithms. The results obtained suggest the method has strong potential as alternative method for building a facial feature tracking system. In the future we hope to include additional features in the tracking.

References

1. Barreto, J., Menezes, P., Dias, J.: Human-robot interaction based on haar-like features and eigenfaces. In: Proceedings of the International Conference on Robotics and Automation, New Orleans, pp. 1888–1893 (2004)
2. Fasel, B., Luettin, J.: Automatic Facial Expression Analysis: A Survey. *Pattern Recognition* 36(1), 259–275 (2003)
3. Tiddeman, B., Perrett, D.: Moving Facial Image Transformations using Static 2D Prototypes. In: Proceedings of the 9-th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision 2001 (WSCG 2001), Plzen, Czech Republic, February 5-9 (2001)
4. Chen, J., Tiddeman, B.: A stereo head pose tracking system. In: Proceedings of the 5th IEEE International Symposium on Signal Processing and Information Technology, Greece, December 18-21, 2005, pp. 258–263 (2005)
5. Yang, J., Stiefelhagen, R., Meier, U., Waibel, A.: Real time face and facial feature tracking and applications. In: Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP 1998, pp. 207–212 (1998)
6. Stiefelhagen, R., Meier, U., Yang, J.: Real-time lip-tracking for lip reading. In: Proceedings of the Eurospeech 1997, 5th European Conference on Speech Communication and Technology, Rhodos, Greece (1997)
7. Tian, Y., Kanade, T., Cohn, J.F.: Recognizing upper face action unit for facial expression analysis. In: Proceedings of the International Conference on Computer Vision and Pattern recognition, South Caroline, USA, June 2000, pp. 294–301 (2000)

8. Kapoor, A., Picard, R.W.: Real-Time, Fully Automatic Upper Facial Feature Tracking. In: Proceedings of the 5th International Conference on Automatic Face and Gesture Recognition, Washinton DC, USA, May 2002, pp. 10–15 (2002)
9. Matsumoto, Y., Zelinsky, A.: An algorithm for real time stereo vision implementation of head pose and gaze direction measurement. In: Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition, France, pp. 499–505 (2000)
10. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(6), 681–685 (2001)
11. Matthews, I., Baker, S.: Active Appearance Models Revisited, Technical report: CMU-RI-TR-03-02, the Robotics Institute Carnegie Mellon University (2002)
12. Cristinacce, D., Cootes, T.: Feature detection and tracking with constrained local models. In: Proceedings of British Machine Vision Conference, UK, pp. 929–938 (2006)
13. Viola, P., Jones, M.: Robust real time object detection. In: Proceedings of the 2nd International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing and Sampling, Vancouver, Canada (July 2001)
14. Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. *International Journal of Computer Vision* 61, 55–79 (2005)
15. Bourel, F., Chibelushi, C.C., Low, A.A.: Robust Facial Feature Tracking. In: Proceedings of the Eleventh British Machine Vision Conference, Bristol, UK (September 2000)
16. Feng, G., Yuen, P.: Multi-cue eye detection on grey intensity image. *Pattern Recognition* 34, 1033–1046 (2001)
17. Kanade, T.: Picture processing by computer complex and recognition of human faces. Technical report, Kyoto University (1973)
18. Peng, K., Chen, L., Ruan, S., Kukharev, G.: A Robust Algorithm for Eye Detection on Gray Intensity Face without Spectacles. *Journal of Computer Science & Technology* 5(3), 127–132 (2005)
19. Lucas, B., Kanade, T.: An interactive image registration technique with an application in stereovision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence, pp. 674–679 (1981)
20. Trucco, E., Verri, A.: Introductory techniques for 3-d computer vision. Prentice-Hall, New Jersey (1998)
21. Zhu, Z., Ji, Q.: 3D Face Pose Tracking From an Uncalibrated Monocular Camera. In: Proceedings of the 17th International Conference on Pattern Recognition, UK, pp. 400–403 (2004)
22. Yao, P., Evans, G., Calway, A.: Using affine correspondence to estimate 3-d facial pose. In: Proceedings of the International Conference on Image Processing, Greece, pp. 919–922 (2001)
23. Fischler, M.A., Bolles, R.C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Comm. of the ACM* 24, 381–395 (1981)
24. Jesorsky, O., Kirchberg, K.J., Frischholz, R.W.: Robust face detection using Hausdorff distance. In: Proceedings of the 3rd International Conference on Audio and Video-based Biometric Person Authentication, Halmstad, Sweden, pp. 90–95 (2001)
25. Hamouz, M., Kittler, J., Kamarainen, J.K., Kalvioinen, H.: Affine invariant face detection and localization using GMM-based feature detectos and enhanced appearance model. In: Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 67–72 (2004)
26. Cristinacce, D., Cootes, T., Scott, I.: A multi-stage approach to facial feature detection. In: Proceedings of British Machine Vision Conference, UK, pp. 277–286 (2004)

Body Language Based Individual Identification in Video Using Gait and Actions

Y. Pratheepan¹, P.H.S Torr², JV Condell¹, and G. Prasad¹

¹ School of Computing and Intelligent Systems, Faculty of Computing and Engineering, University of Ulster at Magee, Northland Rd, Londonderry, N.I.

{p.yogarajah, j.condell, g.prasad}@ulster.ac.uk

² Department of Computing, Oxford Brookes University, Wheatley, Oxford, UK

philiptorr@brookes.ac.uk

Abstract. In intelligent surveillance systems, recognition of humans and their activities is generally the most important task. Two forms of human recognition can be useful: the determination that an object is from the class of humans (which is called human detection), and determination that an object is a particular individual from this class (this is called individual recognition). This paper focuses on the latter problem. For individual recognition, this report considers two different categories. First, individual recognition using “style of walk” i.e. gait and second “style of doing similar actions” in video sequences. The “style of walk” and “style of actions” are proposed as a cue to discriminate between two individuals. The “style of walk” and “style of actions” for each individual is called their “body language” information.

1 Introduction

There is a strong need for smart surveillance systems from security-sensitive areas e.g. banks to alert security officers to a suspicious individual or unknown individual wandering about the premises. Visual systems can deal with and monitor “familiar” and “unfamiliar” human movements. In order to properly interact with people an intelligent system has to detect people and identify them using their body language. Examples of such systems are intelligent security systems that detect unknown people entering restricted areas or interface robots that act as an interface for taking known users commands and presenting results.

The aim of this intelligent surveillance system project is a general framework that groups a number of different computer vision tasks aiming to identify individuals using their “style of walk”, and on the next level to identify individuals through their “style of action”. Gait based individual recognition is explained in Section 2, action based individual recognition is explained in Section 3. Results and conclusions are discussed in Sections 4 and 5 respectively.

2 Identifying Individuals through Their “Style of Walk”

Humans can identify people known to them over long distances, just by their gait. This implies that the motion pattern is characteristic for each person. Motion

information is one of the good cues which can be used to recognize individuals. The most recognized and earliest psychophysical study of human perception of gait was carried out by Johansson [1] using moving light displays (MLD). The initial experiments showed that human observers are remarkably good at perceiving the human motion that generated the MLD stimuli. Cutting et al. [2] studied human perception of gait and their ability to identify individuals using MLD. There has been an explosion of research on individual recognition in recent years but most of them are based on gait [3][4][5]. The above psychological evidence and computational approaches justify that individual recognition can be done using people's walking sequences. i.e. human gait has enough information to discriminate individuals.

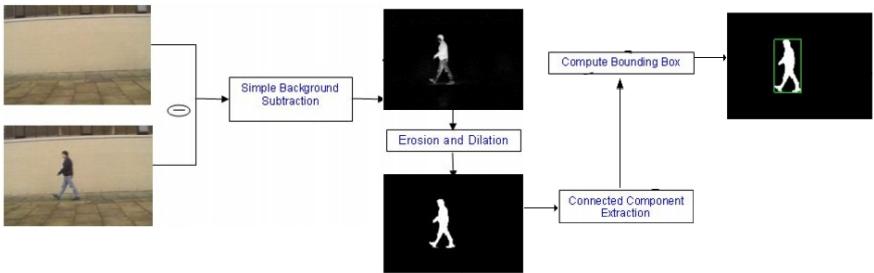
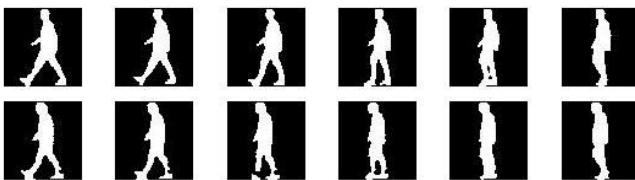
Most of the research has been done using full-cyclic period human walk sequences. The first question is, if it is not possible to obtain full-cyclic walking sequences then how do we recognise individuals?. The number of frames in individuals walk sequences, even from a particular individual's sequence, may vary because of the speed. Therefore the second question is, how do we calculate a similarity measure between the different number of frames based on the "Body-language" information. We attempt to answer these questions in this paper.

2.1 Feature Extraction

This project considers the fronto-parallel view of a walking person to be that which is perpendicular to the direction of walk. The image sequences of four individuals (A, B, C and D) are collected using a stationery camera i.e. walk image sequences with static background. The aim of this work is to recognize individuals using the "style of walk" information alone. That is, the same person can act in walking sequences with different types and colour of clothing, even though the algorithm needs to recognize that individual properly.

To remove the effect of changing clothing colour, only the silhouettes of the walking subjects are used in the representation. In addition, the silhouettes are scaled to remove the effect of changing depth of the walking subject in the view of the camera. For the foreground segmentation the background subtraction is applied and then binarized using a suitable threshold. Morphological operators such as erosion and dilation [6] are first used to further filter spurious pixels. Small holes inside the extracted silhouettes are then filled. A connected component extraction is finally applied to extract a connected region with the largest size motion area of the walking human. The motion area included by the bounding box (Figure 1) is cropped, then re-scaled using a bilinear method [6] into a fixed size $S \times S$ image ($S = 64$, number of pixels). Figure 2 shows normalized silhouettes of two individual's walking sequences.

The size of these cropped images is considerable. Therefore a dimensionality reduction needs to be applied before applying the classification algorithm. This is done using Principal Components Analysis (PCA) on those images.

**Fig. 1.** Silhouette extraction method**Fig. 2.** The silhouettes of individuals A and B's walking sequences

Dimensionality Reduction Using PCA is applied successfully in many applications. Sirovich et al [7] has applied this technique for face recognition. In this work also the PCA technique is used for individual recognition.

The image set is defined as $\{y(i)|i = 1, 2, \dots, M\}$, where M is the number of images in the set. Next the average image \hat{y} is calculated, which is the mean image of all images in the set. An image matrix P is constructed by subtracting \hat{y} from each image and stacking the resulting vectors column-wise. Let, P be $N \times M$, where N is the number of pixels in each image.

In this experiment 114 images are used for training, i.e. $M = 114$. Each image size is 64×64 , i.e. $N = 4096$ and $N > M$. If we consider the covariance matrix Q , where $Q = PP^T$, then Q is $N \times N$ and $N > M$. Calculation of the eigenvectors of a matrix as large as Q is computationally intensive. For improvement the implicit covariance matrix \hat{Q} is used, where: $\hat{Q} = P^T P$.

Note that \hat{Q} is an $M \times M$ matrix and therefore much smaller than Q . Here M eigenvectors of \hat{Q} can be computed. These can be computed much faster than the first M eigenvectors of Q due to the disparity in the size of the two matrices. It can be shown that the M largest eigenvalues and corresponding eigenvectors of Q can be determined from the M eigenvalues and eigenvectors of \hat{Q} as: $\lambda_i = \hat{\lambda}_i$ and $e_i = \hat{\lambda}^{-1/2} P \hat{e}_i$ [8]. Here, λ_i and e_i are the i^{th} eigenvalue and eigenvector of Q , while $\hat{\lambda}_i$ and \hat{e}_i are the i^{th} eigenvalue and eigenvector of \hat{Q} . Previous research has proved that only a few eigenvectors are necessary for visual recognition. Therefore, we use the first k eigenvectors calculated corresponding to the largest

k eigenvalues. The k -dimensional subspace spanned by these eigenvectors is called the *eigenspace*. Singular Value Decomposition (SVD) is applied to the data set as N is much larger than M [8]. It is not viable, however, when more than M eigenvectors are needed.

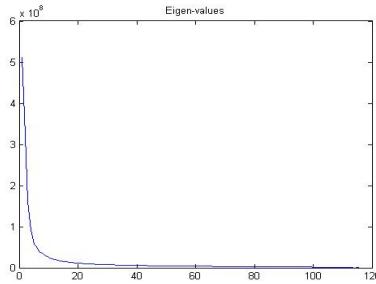


Fig. 3. Eigenvalues obtained from applying SVD algorithm on the entire training sequence

Figure 3 shows the eigenvalues calculated by the SVD algorithm for the training image sequence. The first 40 eigenvalues are greater than zero and then the eigenvalues tend to zero. Therefore, we use the first 40 eigenvectors corresponding to these 40 eigenvalues to reconstruct silhouette sequences. Figure 4 shows the reconstruction result of a silhouette image. Using this dimensionality reduction algorithm, each image in the sequence can be represented by a k dimensional vector and this vector is represented as a **feature vector** for each image.

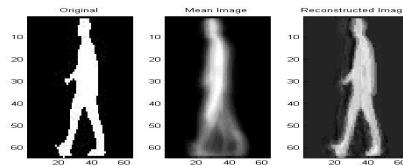


Fig. 4. Left image: Silhouette obtained from test sequence. Middle image: Mean image. Right image: Reconstructed image using first 40 eigenvectors.

An image in the individual sequence can be mapped to a point $f(i)$, where $f(i) = [e_1, e_2, \dots, e_k]^T y(i)$, in the eigenspace. Here, $f(i)$ is the k th dimensional feature vector for image $y(i)$. Therefore, a sequential movement can be represented as a trajectory in the eigenspace. An example of walking patterns for four different individuals is shown in Figure 5. This is the eigenspace representation of the individual trajectory manifolds.

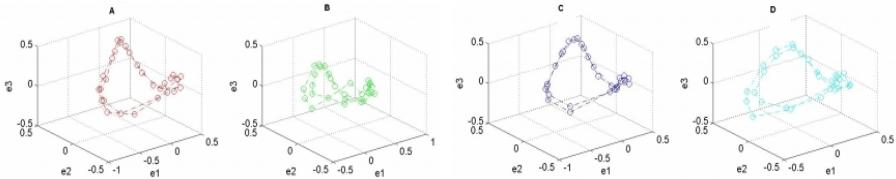


Fig. 5. Trajectory of 4 individuals in 3D space made of first 3 eigenvectors

2.2 Recognition Stage

Walking sequences are periodic and cyclic. In the walking sequence all the phases which match the known poses are called a *walk-cycle*. An example of a human walk-cycle is shown in Figure 6. If an image sequence contains a walk-cycle then that sequence can be called a full walking sequence.

Murase et al [9] and He et al [10] assumed their database had the full cyclic walking sequence data. If any sequence does not have a walk-cycle then we can say those sequences are *partial* sequences. For the partial case the above two methods did not give any solution for partial data comparison. To solve this problem, we needed to find the longest common subsequence from both partial walking sequences. To find the common subsequence, we used the Longest Common Subsequence (LCS) algorithm [11].

In our work we assume walk sequences may consist of full cyclic motion data or partial motion data.

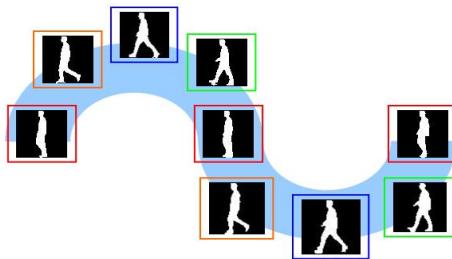


Fig. 6. The silhouettes from a walk-cycle

Longest Common Subsequence (LCS) Algorithm finds the longest subsequence that two sequences have in common, regardless of the length and number of intermittent non-matching symbols. For example, the sequences “**a**c**b**c**d**e**f**g” and “**a**x**b**y**d**e**z**zz” have a sequenced length of four “**a**b**d**e” as their longest common subsequence. Figure 7 given below has more information.

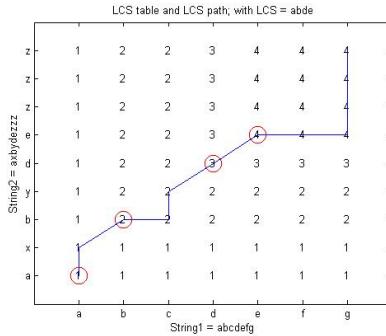


Fig. 7. The longest common sequence “abde” found from two strings “abcdefg” and “axbydezzz” using LCS algorithm

Formally, the LCS problem is defined as follows: Given a sequence $X = (x_1, x_2, \dots, x_m)$, and a sequence $Y = (y_1, y_2, \dots, y_n)$, find a longest common sequence $Z = (z_1, z_2, \dots, z_k)$. The solution to the LCS problem involves solving the following recurrence equation, where the cost for the edit operations stored in C is:

$$C(i, j) = \begin{cases} 0 & \text{if } (i = 0) \text{ or } (j = 0) \\ C(i - 1, j - 1) + 1 & \text{if } (i, j > 0), (x_i = y_j) \\ \max [C(i, j - 1), C(i - 1, j)] & \text{if } (i, j > 0), (x_i \neq y_j) \end{cases} \quad (1)$$

Using LCS as a similarity measure between two sequences has the advantage that the two sequences can have different lengths and have intermittent non-matches. In the context of individual recognition, this allows for the use of partial walking sequences with noisy inputs.

Given two image sequences (S_1 and S_2) the 40-dimensional feature vector is calculated, as described before, for each frame in those sequences. Further we normalize those vectors to unit length to then apply the correlation measure between two frames: $c = x^T y$, where c is the correlation value between 0 and 1, x and y are the normalized vectors of the corresponding frames from S_1 and S_2 respectively. These correlation values are stored in a matrix C . The rows and columns in matrix C represent the frames from sequence S_1 and S_2 respectively. Each value in matrix C tells the degree of similarity between the frames from S_1 and S_2 . From experimental results the correlation value greater than or equal to 0.7 gives similar frames. The threshold value is defined as 0.7 for good experimental results. Now the most similar frames corresponding to both these sequences can be calculated. To do this, we need to find the maximum value for each row. If that maximum value is greater than a threshold value we can say the frames represented by the rows and columns are similar (or equal). It is important to find the similar frames before applying LCS algorithm as there is a calculation in LCS algorithm that, if $(x_i = y_j)$ then $c(i, j) = c(i - 1, j - 1) + 1$. Using the Longest Common Sequence algorithm we can find a set of pair of frames, which

are the similar frames from two walk sequences S_1 and S_2 . The values from the matrix C corresponding to each pair of frames from the set are summed and finally we find the mean value. This mean value gives the final measure of the two sequences.

3 Individual Identification Using “Style of Actions”

The image sequences of three individuals (A, B, and C) are taken. The local space-time intensity gradient (S_x, S_y, S_t) is calculated at each space-time point (x, y, t) for each frame in the sequence. We find the absolute values of the temporal derivatives in all pixels of the image frame and ignore all space-time points (x, y, t) for which the temporal derivative is below some threshold, thus performing the calculation mostly on the points which participate in the event. Let the thresholded image with the detected object be divided into 10×10 bins. Each bin has a numeric value of the total number of pixels which have non-zero values. The values in each bin are stacked into a 100-dimensional vector. This 100-dimensional vector is used as a feature to represent each frame (Figure 8).

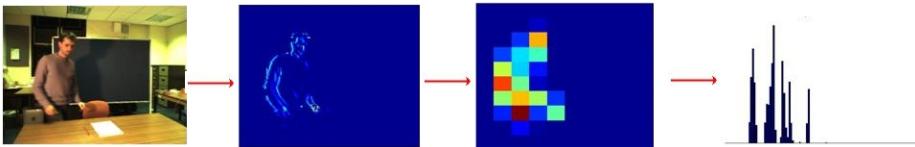


Fig. 8. 100D feature vector generation for image frame in sequence

There are many actions available from the individuals image sequences: enter, pick up the pen, sit down, write their name, get up and leave etc. The frames are clustered into groups to represent the actions (here we assume the number of groups = 9). Consider the whole image frames from these three image sequences and apply the K-means algorithm. After clustering the 100-D vectors in each cluster we represent similar actions. We calculate the mean of the vectors from each cluster and this mean vector is the “code-word”. Each code-word represents a different action.

For the purpose of individual re-identification a similarity measure is needed to match two image sequences, and find whether or not the same individual appears in those sequences. If two image sequences are captured from a particular person in a different time then the length of these two sequences need not be with the equal number of frames depending on both the speed and movement performance. Therefore image sequence alignment is important for the similarity measure. Here the DTW (Dynamic Time Warping) [11] alignment algorithm is used for sequence alignment with an assumption that start and end frames are correctly aligned.

Method: a) Take an image sequence that needs to be matched to another sequence. b) For each frame in the image sequence, find the most suitable code-word. This operation converts an image frame sequence into a code-word sequence. c) Apply the DTW algorithm for code-word sequences to do alignment and find the distance. Actually this alignment is based on actions.

4 Experiments and Results

Using Gait Sequences: In Table 1 the already known individual's data is represented in rows. Frames 1 to 8 are taken from this data as a partial sequence. The column data are taken from the unknown individuals and we need to find which individuals appear in those sequences. Frames from 4 to 13 are taken from this data as partial data. We can expect that the common subsequence should contain frames 4 to 8. Due to noise it varies slightly.

Table 1. Similarity matrix for known - unknown sequences

Individuals	1-Seq	2-Seq	3-Seq	4-Seq
S1	0.9458	0.8305	0.8908	0.8542
S2	0.7586	0.9877	0.8036	0.8006
S3	0.8979	0.8748	0.9571	0.8867
S4	0.8735	0.7285	0.8783	0.9031

For the maximum value in each column it can find the corresponding row (i.e., column represents the unknown person). This maximum value indicates the high similarity between the row and column. Therefore individuals appearing in corresponding rows and columns are the same person. Here also the threshold value is defined as 0.9. Therefore if the highest value is greater than this threshold value then we accept that the same person is available in both sequences.

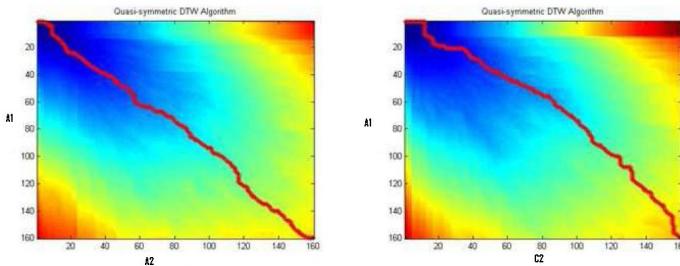
Significant progress has been made in individual identification using their “style of walk” over the past few years. Compared with “style of walk”, not much progress has been made in individuals recognition using their “style of actions”.

Using Action Sequences: For this experiment we have chosen 3 individuals: A, B and C with similar height and width. Table 2 shows the time normalized accumulated distance (see Figure 9) between two sequences. The diagonal elements represent the smallest value in each row. The smallest value shows that individuals appearing in the two sequences are similar based on their “way of doing actions” i.e. “body language”.

We could also apply our LCS based individual recognition technique to this “style of actions” based individual recognition as well.

Table 2. Similarity matrix for known - unknown sequences

Individuals	A1	B1	C1
A2	0.0608	0.2213	0.2976
B2	0.2543	0.0698	0.2588
C2	0.2514	0.1179	0.0917

**Fig. 9.** The path with the minimum distance between two individuals sequences. Path towards the diagonal gives the the minimum distance.

5 Conclusion

This paper has implemented methods for individual recognition based on “*style of walk*” and “*style of actions*”. Both of these methods can be applied to outdoor surveillance e.g. when an unknown individual enters a restricted area and indoor surveillance e.g. when an unknown individual is in an office. The individual recognition method considered the sequence of particular actions (sit down, write name, stand up etc.). It gave good similarity measures between two different individuals actions sequences. To keep the spatial invariance between individuals the individuals were carefully selected with similar height and width. To improve this system we need a larger training set.

Further, to scale up this system into an automated system, as body-language information fully depends on body parts motion, so the body parts must be detected automatically. Optical flow [2] motion detection methods are shown to be more useful for finding particular moving regions e.g. body parts. The main constraint in the DTW method is that the first and last frames in the video sequence should be aligned properly. Ongoing work applies N-Cut clustering to individuals video sequences.

References

1. Johansson, G.: Visual motion perception. *Science American* 232(6), 76–88 (1975)
2. Cutting, J., Prott, D., Kozlowski, L.: A biomechanical invariant for gait perception. *Journal of Experimental Psychology: Human Perception and Performance* 4(3), 357–372 (1978)

3. Little, J.J., Boyd, J.E.: Recognizing people by their gait: The shape of motion. *Computer Vision Research* 1(2), 1–32 (1998)
4. Nixon, M., Carter, J., Nash, J., Huang, P., Cunado, D., Stevenage, S.: Automatic gait recognition. *IEE Colloquium Motion Analysis and Tracking*, 31–36 (1999)
5. Johnson, A., Bobick, A.: Gait recognition using static, activity specific parameters. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 423–430 (2001)
6. Gonzalez, R.C., Woods, R.E.: *Digital Image Processing*, 2nd edn. Pearson Education, London
7. Sirovich,, Kirby, M.: Low dimensional procedure for the charecterization of human faces. *Journal of Optical Society of America* 4(3), 519–524 (1987)
8. Nayar, S.K., Murase, H., Nene, S.A.: Parametric appearance representation in early visual learning, ch. 6. Oxford University Press, Oxford (1996)
9. Murase, H., Sakai, R.: Moving object recognition in eigenspace representation: gait analysis and lip reading. *Pattern Recognition Letters* 17(2), 155–162 (1996)
10. He, Q., Debrunner, C.H.: Individual recognition from periodic activity using hidden markov models. In: *Proceedings IEEE Workshop on Human Motion*, pp. 47–52 (2000)
11. Guo, A., Siegelmann, H.: Time-warped longest common subsequence algorithm for music retrieval. In: *Proc. Fifth International Conference on Music Information Retrieval*, pp. 10–14 (2004)
12. Condell, J.V., Scotney, B.W., Morrow, P.J.: Adaptive Grid Refinement Procedures for Efficient Optical Flow Computation. *International Journal of Computer Vision* (1), 31–54 (2005)

Human Hand Gesture Recognition Using Motion Orientation Histogram for Interaction of Handicapped Persons with Computer

Maryam Vafadar and Alireza Behrad

Faculty of Engineering, Shahed University
Tehran, Iran

vafadar_ma@yahoo.com, behrad@shahed.ac.ir

Abstract. Some groups of handicapped persons cannot reliably move the mouse and do the necessary operation on it to control the computer. However they can do some 3-d hand motions. Variety of tools has been presented for these users to interact with computer. Hand gesture recognition is one of the proper methods for this purpose. This paper presents a new algorithm for hand gesture recognition. In this algorithm, after constructing motion history image of video frames for each gesture and applying necessary processing on this image, motion orientation histogram vector is extracted. These vectors are then used for the training of Hidden Markov Model and hand gesture recognition. We tested the proposed algorithm with different hand gestures and results showed the correct gesture recognition rate of 90 percent. Comparing the results of proposed method with those of other methods showed that in addition to eliminating traditional problems in this area, recognition rate has been improved up to 4 percent.

Keywords: Hand gesture recognition, Motion history image, Hidden Markov model, Motion orientation histogram.

1 Introduction

Some groups of handicapped persons cannot reliably move their hand motions. Although their mental abilities are similar to others, they can not use mouse to interact with computer. However they can do some 3-D hand motions. New tools are needed for these users to interact with computer without mouse limitations [1]. Sensors and video cameras are two sample tools that have been presented for this purpose. Sensor outputs need some filter stages to eliminate temperature, pressure and other noise effects. In addition to low accuracy, higher cost of these devices put them out of reach for general use. Webcams are cheap and widespread [2]. By applying video image processing, they can be used for gesture recognition and interaction of handicapped persons with computer.

Vision-based hand gesture recognition methods are generally categorized in two groups: feature/model based [3] and appearance/view based [4] methods.

In feature/model based methods, initially, we need to extract model or features from images [5]. Implementing this stage often leads to three problems [6]:

1. High computational approximations increase the error rate.
2. High computational complexity slows down the system and makes it unsuitable for real-time interactions.
3. General hand gesture recognition without special features in image frames is not possible.

In appearance/view based methods, images can be used directly for hand gesture recognition. So unlike feature/model based methods, we do not need feature extraction in initial stage [6] and problems of this stage eliminated automatically.

Different appearance/view based methods have been proposed for the task of hand gesture recognition including applying Fourier transform [7], wavelet transform [8] or Principal Component Analysis (PCA) [9-11] on images. Edge orientation histogram also is used for static hand gesture recognition [12]. Using temporal templates is another appearance/view based method for gesture recognition [13-15]. In this method, the basic idea is the projection of the temporal pattern of motion into a single image-based representation called Temporal Template. Then appropriate features are extracted from this image. Temporal Templates have been used for face gesture recognition [13] and hand gesture recognition [14-15].

Several methods have used motion history images (MHI) as temporal templates for gesture recognition [16-19]. In these methods a sequence of image motions are used to constitute a single static image. Then appropriate features are extracted from this image and different methods are used for classification such as least Mahalanobis distance [18], least Euclidian distance [19], Recurrent Neural Network (RNN) [15] and Hidden Markov Model (HMM) [20].

This paper presents a new algorithm for hand gesture recognition. In this algorithm, after constructing motion history image of video frames for each gesture and applying necessary processing on it, motion orientation histogram vectors are extracted. These vectors are then used for the training of Hidden Markov Model and hand gesture recognition. We tested the proposed algorithm with the collected data set and results showed the correct gesture recognition rate of 90 percent.

The paper continues as follows: In section 2, we describe the collected data set of hand gestures. Section 3 explains the method of constructing MHI (Motion History Image). Section 4 presents the extraction of motion orientation histogram from MHI. In section 5, gesture classification using HMM (Hidden Markov Model) is presented. Section 6 shows experimental results for the proposed method and conclusions appears in section 7.

2 Collected Data Set

Data set is collected from 5 persons with different hand size and skin color. No obvious feature or sign exists on the hands of these persons. The collected data set contains different videos of six hand gestures. These gestures are 6 different motions of the hand including right, left, up, downward motion and closing and opening of the

fingers, which is denoted as a, b, c, d, e and f gestures, respectively. Gestures a-d may simulate mouse right, left, up and downward motions and two other motions can be used for the simulation of mouse click. We used 5 different persons for tests and gesture was captured in different conditions. Each person repeated these gestures 10 times in different illumination condition and with different backgrounds. Since handicapped persons may not move their hands exactly in one plane, some of these gestures are not intentionally in a plane. Resolution of the video frames is 176*144.

3 Constructing MHI

3.1 Motion History Image (MHI)

Motion History Image (MHI) is a single static image which its pixel intensity is the function of the motion history of captured sequence at that point. MHI is constructed by successively layering the selected image regions over time. Equation (1) presents constructing MHI.

$$H_\tau(x, y, t) = \begin{cases} \tau & \text{if } D(x, y, t) = 1 \\ \max(0, H_\tau(x, y, t-1) - 1) & \text{otherwise} \end{cases} . \quad (1)$$

where H_τ is Motion History Image(MHI) and D is the binary difference image that we will explain it in section 3.2. x and y are image pixels coordinates and t is temporal index. τ , is a threshold for extraction moving patterns in video image sequence. Thus, MHI is a scalar-valued image where more recently moving pixels are brighter.

3.2 Difference Image

Depending on MHI applications, different methods may be used for constructing difference image. One method is the construction of background image and subtraction of the frames from background image [19]. In our application, we may have no information about the background [21]; therefore we used the difference of consecutive frames as difference image after applying a threshold.

Figure 1 shows an example of Motion History Image constructed from difference image. As it is obvious from this figure, more recently moving pixels are brighter than other pixels. Indeed, MHIs show the direction of motion, and we can use this feature for gesture recognition in next sections.

3.3 Noise Filtering

Any noise in video frames is directly transferred to difference images and appears in MHI. This may degrade the performance of gesture recognition algorithm; therefore it is necessary to handle noise.

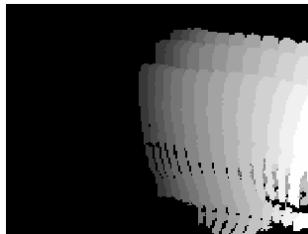


Fig. 1. An example Motion History Image for rightward hand motion

We used morphology operators to remove noise. We applied proper morphology operators to binary image to remove connected component with the size of smaller than 20 pixels. Our experimental results showed that $P=20$ is an optimized value which removes the noise effects and preserves motion information properly. Figure 2 shows the result of noise removing algorithm.

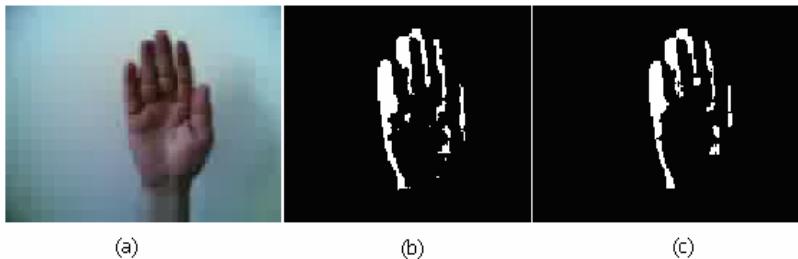


Fig. 2. (a) First frame, (b) difference image and (c) filtered image

3.4 Hand Size Scaling

Proposed algorithm must be insensitive to hand sizes and several users with different hand sizes may use this method. Different hand sizes-even for a single gesture- may have different Motion History Images and results in erroneous results which is not desirable. We applied a normalization algorithm to handle this problem. In normalization methods, hands with different sizes are projected to a constant size. Different algorithm may be used for normalization. In [5] special point are extracted from input image, and distance and angle of the selected points are calculated with respect to image center. These parameters are then used to calculate scaling factor and normalization. Our proposed algorithm doesn't need to extract special points of the image. In this method, we extract the contour of hand motion in MHI using edge detection algorithm. Then rectangular bounding box of the contour is used to constitute an image containing only moving part of the MHI. The resultant image is then resized to standard size of 88*72. This image is used for hand gesture recognition in next sections. Figure 3 shows the result of normalization algorithm.

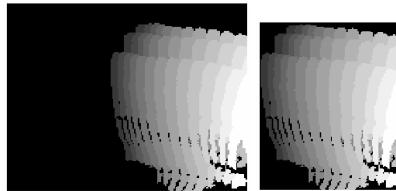


Fig. 3. Hand size normalization of MHI

4 Motion Orientation Histogram

According to equation (1) hand motion results in intensity variation in MHI. So using gradient operator, we can extract motion vector in each pixel and calculate motion fields. We applied Sobel operator to resultant image of previous stage to calculate motion fields as follows:

$$I_x = I * G_x . \quad (2)$$

$$I_y = I * G_y . \quad (3)$$

$$\theta(i, j) = \tan^{-1} \frac{I_y(i, j)}{I_x(i, j)} . \quad (4)$$

where I is Motion History Image of previous stages, G_x and G_y , are Sobel gradient operators. I_x and I_y , are image gradients in x and y directions, showing motion fields and θ is motion orientation and i, j are image pixel coordinates.

When the motion fields are calculated, they are used to extract proper features for classification and gesture recognition. We use the histogram of motion orientations as features for classification and gesture recognition. The features are insensitive to the velocity of motion; therefore they can be used for recognition of the same gesture with different velocities. Figure 4 shows an example of hand motion orientation histogram. To calculate motion orientation histogram, motion orientation is calculated

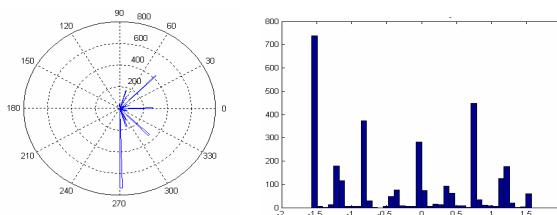


Fig. 4. The orientation histogram for rightward hand motion

using equation (4), then orientation image is divided to 40 bins in polar coordinate, then orientation of points in each bin is summed to obtain orientation histogram. So motion orientation histograms are one dimensional vector of length 40, which are used for classification and gesture recognition.

5 Classification

We use Hidden Markov Model to classify 6 hand gestures. We used left to right model of HMM for classification. This model is appropriate for dynamic gesture recognition. We use One HMM for each hand gesture. In each HMM, motion orientation histogram vectors determines system states. Using training set of motion orientation histogram and baum-welch algorithm, probability parameters are calculated. For hand gesture recognition, after extracting motion orientation histogram, output probabilities $P(O | \lambda^g)$ for all trained models are computed by Viterbi algorithm. At last, a model with highest probability is selected as output.

It is important to note that, we extract orientation histogram for all video samples of a gesture and train all of them to HMM model of that gesture. Weighted Gaussian functions are generated for each gesture by applying k-means algorithm.

6 Experimental Results

We tested the proposed algorithm with our collected data set. We used different combination of gestures as training sets and test inputs. Figure 5 shows hand motion orientation histogram for 6 different gestures. As it is shown in this figure, histograms are different.

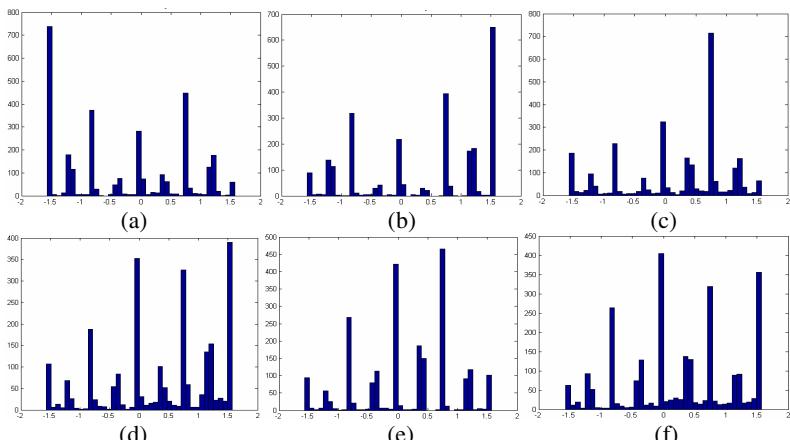


Fig. 5. Motion orientation histograms of (a) rightward (b) leftward (c) upward (d) downward (e) closing of fingers and (f) opening of fingers

We totally conducted 50 different experiments for testing the proposed algorithm and comparing the results with those of other methods. Table 1 shows the results of gesture recognition algorithm for 50 experiments.

To compare the results of proposed method with other methods, we also implemented the algorithms of [18] and [19]. Table 2 shows recognition rates for the proposed and methods of [18] and [19]. Results show the correct gesture recognition rate of 90 percent for the proposed method. Comparing the results of proposed method with two other methods shows that the recognition rate has been improved up to 10 and 4 percent with respect other methods.

Table 1. The result of gesture recognition using the proposed algorithm

<i>f</i>	e	d	c	b	a	Input gestures	Algorithm recognition
1	0	0	0	0	47		a
0	1	1	1	48	0		b
0	0	0	45	1	1		c
3	4	44	0	0	0		d
2	42	3	1	1	1		e
44	3	2	3	0	1		f

Table 2. Hand gesture recognition results

method	First method [18]	Second method [19]	Proposed method
Recognition Rate	80,3%	86%	90%

7 Conclusion

A new algorithm based on Motion History Image for hand gesture recognition is presented. In this algorithm, after constructing motion history image of video frames for each gesture and applying necessary processing on this image, motion orientation histogram vector is extracted. These vectors are then used for training of Hidden Markov Model and hand gesture recognition. We tested the proposed algorithm with the collected data set and results showed the correct gesture recognition rate of 90 percent. Comparing the results of proposed method with those of other methods showed that recognition rate has been improved up to 4 percent. In addition the

proposed algorithms is robust to traditional problems of gesture recognition like illumination variations, different skin color and hand sizes. The algorithm can be used for interaction handicapped persons with computer and increase their abilities.

References

1. Derpanis., G.: A Review of Vision-Based Hand Gestures. York University (2004)
2. Hai., W., Shamaie., A., Sutherland., A.: Real Time Gesture Recognition of Human Hand <http://www.computing.dcu.ie/~alistair/CA107.ppt>
3. Mu-Chun, S.: A Neural Network-Based Approach to Recognizing 3D Arm Movement, Biomedical Engineering Application. Basis and Communication 15(1), 17–26 (2003)
4. Hou., X., Li., S.Z., Zhang., H., Cheng., Q.: Direct Appearance Models. In: Proc. Of IEEE Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 828–833 (2001)
5. Shalbaf, R., Vafadoost, M., Shalbaf, A.: Lipreading Using Image Processing for Helping Handicap. In: Iranian conf. on Biomedical Engineering (2007)
6. Trucco., E., Verri., A.: Introductory Techniques for 3-D Computer Vision. Prentice-Hall, New Jersey (1998)
7. Harding., P.R.G., Ellis., T.: Recognition Hand Gesture Using Fourier Descriptors. In: Proc. of IEEE Conf. on Pattern Recognition (ICPR 2004), vol. 3, pp. 286–289 (2004)
8. Dionisio, C.R.P., Cesar Jr., R.M.: A Project for Hand Gesture Recognition. In: Proc. of IEEE Symposium on Computer Graphics and Image Processing, p. 345 (2000)
9. Lamar., M.V., Bhuiyan, M.S., Iwata., A.: Hand Gesture Recognition Analysis and an Improved CombNET-II. In: Proc. Of IEEE Conf. on Man and Cybernetics, vol. 4 (1999)
10. Moghaddam, B., Pentland, A.: Probabilistic Visual Learning for Object Detection. In: Conf. on Computer Vision, Cambridge, MA (1995)
11. Moghaddam, B.: Principal Manifolds and Bayesian Subspaces for Visual Recognition. In: Proc. Of IEEE Conf. on Computer Vision, ICCV 1999 (1999)
12. Freeman, W.T., Roth, M.: Orientation Histogram for Hand Gesture Recognition. In: IEEE Int. Workshop on Automatic face and gesture recognition (1995)
13. Black, M., Yacoob, Y.: Tracking and Recognizing Rigid and Non-rigid Facial Motion Using Local Parametric Models of Image Motion. In: Proc. Of IEEE Conf. on Computer Vision, ICCV 1995 (1995)
14. Shan., C., Yucheng., W., Xianchao., Q., Tieniu., T.: Gesture Recognition Using Temporal Template Based Trajectories. In: Int. Conf. on Pattern Recognition, vol. 3, pp. 954–957 (2004)
15. Kumar, S., Kumar, D.K., Sharma, A., McLachlan, N.: Classification of Hand Movements Using Motion Templates and Geometrical based Moments, 3, pp. 299–304 (2004)
16. Venkatesh Babu, R., Ramakrishnan, K.R.: Recognition of Human Actions Using Motion History Information Extracted from the Compressed Video, vol. 22, pp. 597–607 (2004)
17. Ogata, T., Tan, J.K., Ishikawa, S.: High-Speed Human Motion Recognition Based on a Motion History Image and an Eigenspace. IEICE Trans. on Information and Systems E89-D, 281–289 (2006)
18. Bobick., A.F.: Computers Seeing Action, MIT media library (1997)
19. Davis., J.W.: Hierarchical Motion History Images for Recognizing Human Motion, Department of computer and Information science of Ohio state university (2002)
20. Kenny, M., McKenna, S.J.: An Experimental Comparison of Trajectory-Based and History-Based Representation for Gesture. In: Camurri, A., Volpe, G. (eds.) GW 2003. LNCS (LNAI), vol. 2915, pp. 152–163. Springer, Heidelberg (2004)
21. Keskin, C., Erkan, A., Akarun, L.: Real time Hand Tracking and 3D Gesture Recognition for Interactive Interfaces using HMM, Bogazici university, Eusipco (2005)

3D Human Motion Reconstruction Using Video Processing

Nadiya Roodsarabi and Alireza Behrad

Faculty of Engineering, Shahed University

Tehran, Iran

n_rsr20@yahoo.com, behrad@shahed.ac.ir

Abstract. One of the important problems in human motion analysis is the 3D reconstruction of human motion, which utilizes the anatomic point's positions. These points can uniquely define the position and orientation of all anatomical segments. In this paper, a new method for reconstruction of human motion from the image sequence of a single static camera is described. In this method 2D tracking is used for 3D reconstruction, which a database of selected frames are used for the correction of tracking process. We use Discrete Cosine Transform (DCT) block as a “Matrix des criptor” used in the matching process for finding appropriate frame in the database and tracking process. Finally, 3D reconstruction is performed using Taylor method. By using DCT, we can select best frequency region for various tasks such as tracking, matching, correcting joints and so on. Experimental results showed the promise of the algorithm.

Keywords: Discrete Cosine Transform (DCT), Human motion reconstruction, Video processing, Occluded limb, Pose corresponding, Tracking, Matching.

1 Introduction

The realistic modeling and animation of human characters is one of the most difficult tasks in the vision and graphic community. In particular, body modeling from video sequences is a challenging problem that has been investigated a lot in the last decade. Recently the demand of 3D human models is drastically increased for applications like movies, video games, ergonomic, e-commerce, virtual environments and medicine.

A classical approach to build human shape models uses 3D scanners [1]. They are some disadvantages such as expensive instrument, limited flexibility (heavy wires) and freedom constraints, but they are simple to use and various softwarse are available to model the resultant measurements.

Due to the high number of degrees of freedom of the human body, motion tracking is a difficult problem. The problem is particularly challenging for visionbased (marker-less) approaches because of self occlusion, unknown kinematic information, perspective distortion and cluttered environments.

The existing approaches to human motion analysis can be roughly divided into two categories: model-based methods and model-free methods. In the model-based methods, a priori human model is used to represent the observed subjects, while in model-free methods; the motion information is derived directly from a sequence of

images. The main drawback of model-free methods is that they are usually designed to work with images taken from a known viewpoint. Model-based approaches, on the other hand, support viewpoint independent processing and have the potential to generalize across multiple viewpoints.

One decisive factor determining the used methodologies is the input data characteristics provided to the system by the acquisition stage. Some authors use monocular views [2-5], while others focus on multi-camera video streams [6-10]. Some limit their work to the use of calibrated views [8,9],[11-13], while others choose noncalibrated images [2]. Extracting monocular 3D human motion poses a number of difficulties such as Depth 3D-2D Projection Ambiguities, High-Dimensional Representation, Physical Constraints, Self-Occlusion and Observation Ambiguities.

Nowadays, monocular uncalibrated video sequences are the most common source of human motions. These methods which enable the extraction of detailed information about a specific uncalibrated sequence would greatly benefit applications related to video compression, video content-based classification and annotation industries. At a different level, the idea of being able to recover the motion of actors or historical celebrities from old movies and bringing them to life in new movies, animations, games or virtual environments is also very attractive. To compensate the lack of calibration, manual specification of key features such as joints or adjustment of a reference skeleton to specific frames are crucial [14]. To facilitate the process, motion databases can be used, becoming indispensable the containment of similar motion clips to the motion being recovered [15, 16].

Monocular methods for motion reconstruction are divided in two categories: 1-discriminative methods [4-5] 2- estimating and tracking methods [3]. In deiscriminative methods, 3D joint coordinates are found by using database, motion libraries and so on. In estimating and tracking methods, 3D information at a step is estimated using a sequence of images, prior and posterior to it.

Table 1. Tasks and required frequency regions

Tasks	Low frequency	Middle frequency	High frequency
Used frequency region for matching process		×	
used frequency region for database matching process	×	×	
Check tracking errors		×	
Updating RDMs if true tracking	×	×	×
Updating RDMs if false tracking			×
Updating RDMs if database matching occur		×	

In this article, we introduce a new descriptor with Discrete Cosine Transform used to various tasks in the algorithm. Tracking and matching is based on Reference

Matrix Descriptors (RMDs). These RDMs must be updated after each stage based on the frequency regions. Advantage of using descriptor is the capability of selecting required frequency in various tasks which results in better tracking and pose matching. For example, we use low and middle frequency in tracking for intensity and edge tracking. Also, we pass up color of clothes in database matching with avoiding low frequency.

Table 1 shows all tasks and their required frequency regions.

Organization: The paper is organized as follows. In section 2 we review the model considered to model 3D human motion. In Section 3 we give an overview of the algorithm of reconstructing 3D human motion using sequences of images acquired with a single video camera. Finally, we report the practical results in section 4.

2 Human Body Model

Human skeleton system is treated as a series of jointed links (segments), which can be modeled as rigid bodies. In the motion reconstruction applications, it is common to use a simple skeleton system modeling the important segments. We describe the body as a stick model consisting of a set of thirteen joint (plus the head) connected by thirteen segments as shown in Table 2.

Table 2. Relative lengths of the segments

Segment	Relative Length (MC) [cm]	Relative Length (L) [unit]
Height	175	8 i
Lower arm	35	2 i
Upper Arm	25	1 ½ i
Neck-Head	25	1 ¼ i
Shoulder Girdle	44	2 i
Torso	53	2 ½ i
Pelvic Girdle	30	1 ½ i
Upper leg	46	2 i
Lower leg	52	2 i
Foot	22	1 i

The algorithm needs the knowledge of relative lengths of the segments for the 3D reconstruction purpose, which can be obtained from anthropometric data. Table 2 shows this relative length. With known 2D position and using the knowledge of length of the segments and enforcing some constraints such as dynamic smoothing, we can obtain 3D reconstruction.

3 Overview of Algorithm

In this method, we locate 2D joints position using a fixed and uncalibrated monocular video and use them to estimate 3D skeletal configuration. As regards no enough information is available from monocular video; we save several 2D exemplar of various body poses in the database that used to correct tracked points.

In the proposed algorithm, joint tracking is performed based on a $n \times n$ block of DCTs coefficient (descriptor matrix). Algorithm starts with background subtraction and is initialized by user through specifying 2D joint positions in first frame. Then for each joint, descriptor matrix is computed and saved as “Reference Descriptor Matrix”

for the same joint. In the next stage, all joints are tracked with its own RDM. After finding joint positions in next frame, RDMs are updated based on DCT block frequency regions considering occlusion problem and tracking errors. Then, human pose in current frame is compared with poses in database based on middle frequency. If corresponding occurred, joint positions are corrected and RDMs are updated. The reason for using middle frequency is that we want pass up clothing color (low frequency) and body deformation details (high frequency).

A major problem may be encountered is the occlusion of joints. We solve this problem with “occluded” label for each tracked joints. After tracking process for all frames, “occluded” labeled joints are corrected by interpolation. Given the 2D joint locations, the 3D body configuration and pose are estimated using the algorithm of Taylor [18]. Figure 1 shows the overview of algorithm.

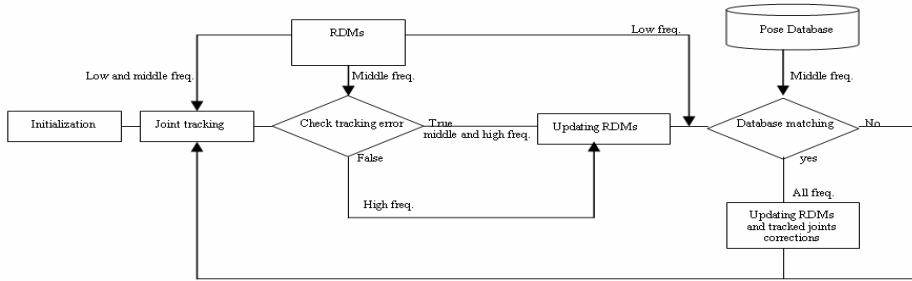


Fig. 1. Overview of algorithm

3.1 Descriptor Matrix

In this article, we use DCT block for tracking and matching purposes. “Descriptor Matrix” for the point p_i is $n \times n$ DCT coefficient matrix. By putting the image window of fixed size ($n \times n$) centered at point p_i into matrix A, a descriptor matrix (DM) for p_i is then compute by equation 1.

$$F(u, v) = C(u)C(v) \left[\sum_{x=px-\frac{n}{2}}^{x=px+\frac{n}{2}} \sum_{y=py-\frac{n}{2}}^{y=py+\frac{n}{2}} f_i(x, y) * \cos \frac{(2x+1)u\pi}{2n} \cos \frac{(2y+1)v\pi}{2n} \right] \quad (1)$$

where px and py are p_i coordinate. Also, if $x=0$; $C(x)=1/\sqrt{n}$ otherwise $C(x)=\sqrt{2/n}$. There are n^2 coefficients in each DM matrix divided into three frequency regions according to figure 2. White region is low frequency region, gray region is middle frequency region and black region is high frequency region. We use this three frequency region for tracking and matching algorithm.

Matrix distance in special frequency region is defined according to:

$$Mdis_{frequencyregion}(M, N) = \sqrt{\sum_f (M_f^2 - N_f^2)} \quad f \in frequency \quad region \quad (2)$$

DC	1	5	6	14	15	27	28
2	4	7	13	16	26	29	42
3	8	12	17	25	30	41	43
9	11	18	24	31	40	44	53
10	19	23	32	39	45	52	54
20	22	33	38	46	51	55	60
21	34	37	47	50	56	59	61
35	36	48	49	57	58	62	63

Fig. 2. Discrete Cosine Transform coeffitients and frequency regions for a 8*8 block

3.2 Reference Descriptor Matrix (RDM)

These matrices save tracked joints information using previous frames and database information and are used to find corresponding joints in tracking process. We generate a reference descriptor matrix for each joint ($RDM_1 \dots RDM_{13}$). Reference descriptor matrix for joint j (RDM_j) are loaded from descriptor matrix for joint j after initialization and updated after finding new tracked joint j in next frame. Updating routine is different in each frequency region:

Low frequency region: This region consists of general shape and intensity information of tracked joint. So its changes are small in successive frames. Tracking process may lose tracked object for several reasons such as occluding problem or large distortion and tracked joint information may be incorrect. For safekeeping of general object information, we leave the low frequency coeffitients unchanged after tracking. These matrices will be updated only when corresponding to database occure.

Middle frequency region: This region consists of general edge information. Because the individual limbs are deformable due to moving muscle and clothing, we apply changes in middle frequency coeffitients after tracking if tracking is accurate. In order to determine the accuracy of the tracked joint j at frame $t+1$, we calculate matrix distance in middle frequency between descriptor matrix of tracked point ($DM_j(t+1)$) and RDM_j .

$$Mdis_{middle}(RDM_j, DM_j(t+1)) \begin{cases} < \Delta & j(t+1) \text{ is true} \\ > \Delta & j(t+1) \text{ is false} \end{cases} \quad (3)$$

We update this frequency region if $j(t+1)$ is true according to equation 3 after tracking. Also this region is updated when corresponding to database occure.

High frequency region: this region is consists of noise and object details and must be updated after each tracking phase.

3.3 Tracking

The tracking process is based on frequency domain matching techniques.

Tracking process aims to find body joints in successive frames. Because of temporal correspondences between subsequent frames, search for corresponding joint is local. In two successive frames, limbs and joints have the same intensity and general shape, but they are different in details. So we use low and middle frequency in tracking process.

The tracking process is based on DCT matching techniques. Its basic idea is to track joints through the sequence. Descriptor matrices are computed for every pixel in the search window (SWDMs). Finding best match is performed by selecting minimum matrix distance between low and middle frequency of RDM_j and search window descriptor matrices (SWDMs).

Assuming a first estimate of the pose has been given, the tracking algorithm can be summarized in 2 steps as follows:

- 1- Generate descriptor matrices for all pixels in search window at frame $t+1$ (SWDMs)
- 2- Determine best matching point in search window by computing matrix distance between RDM_j and SWDMs.

3.4 Database Matching Process

The database consists of different poses required information of video sequences of number of subjects. This information is body joint positions and middle frequency of their descriptor matrices and necessary labels for 3D reconstruction. Head position has been selected as reference joint to adjust two poses and other joint positions have determined towards it.

Measuring similarity between human pose in current frame (p_f) and human pose in database (p_d) requires two kinds of matrix: DDMs and FDMs, which will be defined later. If pose distance is smaller than Δ , corresponding occurs; therefore points and middle frequency of RDM must be corrected. Pose distance is defined by:

$$Pdis_p(p_f, p_d) = \sum_{j=1}^{13} Mdis_{low,mid}(DDM_j, FDM_j) \quad (4)$$

Database descriptor matrix (DDM) is generated using low frequency of RDM (for intensity joint similarity) and midlle frequency of database (for edge similarity).

$$DDM_f = \begin{cases} RDM_f & f : low \text{ frequency} \\ Database & f : middle \text{ frequency} \\ 0 & f : high \text{ frequency} \end{cases} \quad (5)$$

Frame descriptor matrix (FDM) is generated by following process:

- 1- Search locally around the previous head position to find corresponding to RDM_{head} point in the current frame.
- 2- Determine other joints in current frame with adjusting head position.
- 3- Generate descriptor matrices for each joint and save them as FDMs.

Measuring similarity algorithm between human pose in current frame (p_f) and human pose in database (p_d) can be summarized as follows:

- 1- Generate DDMs.
- 2- Search locally to find head position in current frame.
- 3- Determine other point positions in the current frame.
- 4- Compute matrix distance for DDM and FDM in low and middle frequency.
- 5- Correct points if corresponding occur.
- 6- Updating RDMs.

3.5 Estimating 3D Reconstruction

We use Taylor's method [18] to estimate the 3D configuration of a body given the keypoint position estimates. Taylor's method works on a single 2D image, taken with an uncalibrated camera.

It assumes that we know:

- 1- the image coordinates of keypoints (u, v).
- 2- the relative lengths l of body segments connecting these keypoints.
- 3- a labelling of "closer endpoint" for each of these body segments.
- 4- that we are using a scaled orthographic projection model for the camera.

In our work, the image coordinates of keypoints are obtained via the deformable matching process. The "closer endpoint" labels are supplied on the exemplars, and automatically transferred to an input image after the matching process. The relative lengths of body segments are fixed in advance, but could also be transferred from exemplars. We use the same 3D kinematic model defined over keypoints as that in Taylor's work.

We can solve for the 3D configuration of the body $\{(X_i, Y_i, Z_i) : i \in \text{keypoints}\}$ up to some ambiguity in scale s . The method considers the foreshortening of each body segment to construct the estimate of body configuration. For each pair of body segment endpoints, we have the following equations:

$$l^2 = (X_1 - X_2)^2 + (Y_1 - Y_2)^2 + (Z_1 - Z_2)^2 \quad (6)$$

$$(u_1 - u_2) = s(X_1 - X_2) \quad (7)$$

$$(v_1 - v_2) = s(Y_1 - Y_2) \quad (8)$$

$$dZ = (Z_1 - Z_2) \quad (9)$$

$$dZ = \sqrt{l^2 - ((u_1 - u_2)^2 + (v_1 - v_2)^2)} / s^2 \quad (10)$$

To estimate the configuration of a body, we first fix one keypoint as the reference point and then compute the positions of the others with respect to the reference point. Since we are using a scaled orthographic projection model the X and Y coordinates are known up to the scale s . All that remains is to compute relative depths of endpoints dZ . We compute the amount of foreshortening, and use the user-supplied "closer endpoint" labels from the closest matching exemplar to solve for the relative depths.

Moreover, Taylor notes that the minimum scale s_{\min} can be estimated from the fact that dZ cannot be complex.

$$s \geq \frac{\sqrt{(u_1 - u_2)^2 + (v_1 - v_2)^2}}{l^2} \quad (11)$$

This minimum value is a good estimate for the scale since one of the body segments is often perpendicular to the viewing direction.

4 Experimental Results

The proposed algorithm is applied for reconstructiong of human subjects from single-camera video. The database consists of some poses of number of subjects, performing different types of motions from the CMU MoBo Database. On this collection of poses, we manually determined joint locations of each pose and “closer endpoint” labels for each body segment used in Taylor’s method. Also, we save middle frequency of descriptor matrix for each labeled joint.

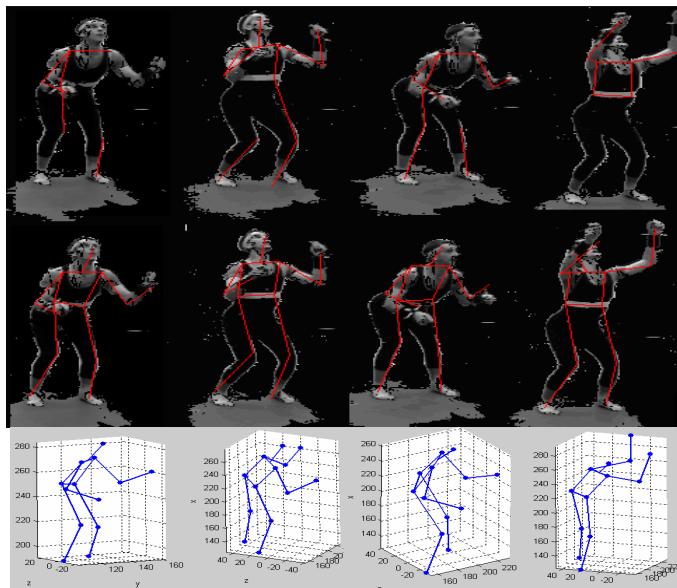


Fig. 3. Reconstruction results for 4 frames. Row 1 shows 2D tracking before interpolation for some frames of the video sequence (7, 14, 54 and 79). Row 2 shows 2D tracking after interpolation. Row 3 shows 3D reconstruction results.

Background subtraction was used to facilitate tracking. We use a skeleton with 13 joints in our experiment and apply a 16*16 DCT block as a descriptor for each point. Then 2D and 3D reconstruction is performed. Figure 3 shows sample results of 2D body joint localization before and after interpolation and finally 3D reconstruction on the CMU dataset. Note that some joints are occluded or failed in 2D tracking. These joints are reconstructed by interpolation. The same body parameters (lengths of body segments) are used in all 3D reconstructions. Figure 4 shows a comparison between 2D optical flow tracking using iterative Lucas-Kanade method in pyramids [20] and our 2D tracking method. As it is shown our method has better results.

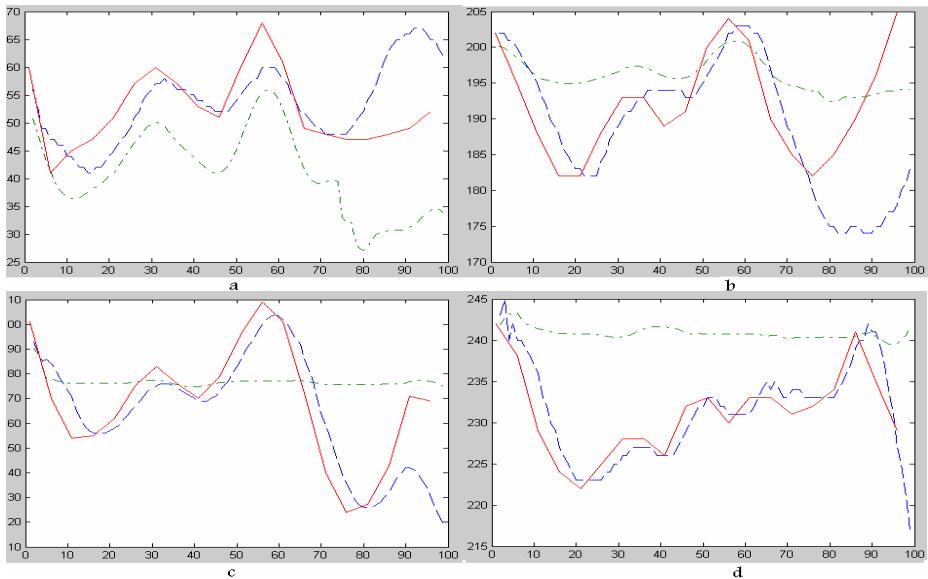


Fig. 4. Comparison between our tracking method and optical flow method using iterative Lucas-Kanade method in pyramids. a) True head position along the x axis (—), our method (- - -) and optical flow method (-.-.). B) True head position along the y axis (—), our method (- - -) and optical flow method (-.-.). c) True right hand position along the x axis (—), our method (- - -) and optical flow method (-.-.). d) True right hand position along the y axis (—), our method (- - -) and optical flow method (-.-.).

5 Conclusion

In this paper, a new method for reconstruction of human motion from the image sequence of a single static camera is described. In this method, 2D tracking is used for 3D reconstruction, which a database of selected frames are used for the correction of tracking process. We used Discrete Cosine Transform (DCT) block as a “Matrix descriptor” used in the matching process for finding appropriate frame in the database and tracking process.

The reconstruction algorithm was tested with several sequences and experimental results showed the reliability of our algorithm. This method is robust in 2D tracking and holding properties of each joint along tracking process.

References

1. Horiguchi: Body Line Scanner.The development of a new 3-D measurement and Reconstruction system. In: International Archives of Photogrammetry and Remote Sensing, vol. 32, pp. 421-429 (1998)
2. Barrón, C., Kakadiaris, I.A.: Estimating anthropometry and pose from single uncalibrated image. Computer Vision and Image Understanding 81(3), 269–284 (2001)

3. Sminchisescu, C., Triggs, B.: Kinematic Jump Processes for Monocular 3D Human Tracking. In: IEEE International Conference on Computer Vision and Pattern Recognition (2003)
4. Chen, C., Zhuang, Y., Xiao, J.: Towards Robust 3D Reconstruction of Human Motion from Monocular Video. In: Pan, Z., Cheok, A.D., Haller, M., Lau, R.W.H., Saito, H., Liang, R. (eds.) ICAT 2006. LNCS, vol. 4282, pp. 594–603. Springer, Heidelberg (2006)
5. Loy, G., Eriksson, M., Sullivan, J., Carlsson, S.: Monocular 3D Reconstruction of Human Motion in Long Action Sequences. In: Pajdla, T., Matas, J.(G.) (eds.) ECCV 2004. LNCS, vol. 3024, pp. 442–455. Springer, Heidelberg (2004)
6. Hilton, A., Beresford, D., Gentils, T., Smith, R., Sun, W., Illingworth, J.: Whole-body modelling of people from multiview images to populate virtual worlds. *The Visual Computer, International Journal of Computer Graphics* 16, 411–436 (2000)
7. D'Apuzzo, N., Plänkers, R., Fua, P., Gruen, A., Thalmann, D.: Modeling human bodies from video sequences. In: El-Hakim, S.F., Gruen, A. (eds.) SPIE. Videometrics VI, vol. 3461, pp. 36–47 (1999)
8. Plänkers, R., Fua, P.: Tracking and modeling people in video sequences. *Computer Vision and Image Understanding* 81(3), 285–302 (2001)
9. Mikić, I., Triverdi, M., Hunter, E., Cosman, P.: Articulated body posture estimation from multi-camera voxel data. In: IEEE Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, vol. I, pp. 455–460 (2001)
10. Cheung, K.M., Kanade, T., Bouguet, J.Y., Holler, M.: A real time system for robust 3D voxel reconstruction of human motions. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 714–720 (2000)
11. Iwasawa, S., Ohya, J., Takahashi, K., Sakaguchi, T., Ebihara, K., Morishima, S.: Human body postures from trinocular camera images. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 326–331 (2000)
12. Plänkers, R., Fua, P., D'Apuzzo, N.: Automated body modeling from video sequences. In: International Conf. of Computer Vision (Workshop on Modeling People), pp. 45–52 (1999)
13. Rosales, R.E., Sclaroff, S.: Learning and synthesizing human body motion and posture. In: IEEE International Conf. on Automatic Face and Gesture Recognition, pp. 506–511 (2000)
14. Park, M.J., Choi, M.G., Shin, S.Y.: Human motion reconstruction from inter-frame feature correspondences of a single video stream using a motion library. In: ACM SIGGRAPH Symposium on Computer Animation, pp. 113–120 (2002)
15. Mori, G., Malik, J.: Estimating human body configurations using shape context matching. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2352, pp. 666–680. Springer, Heidelberg (2002)
16. Park, M.J., Choi, M.G., Gawa, Y.S., Shin, S.Y.: Video-Guided Motion Synthesis Using Example Motions. *ACM Transactions on Graphics* 25(4) (2006)
17. Mori, G., Belongie, S., Malik, J.: Shape contexts enable efficient retrieval of similar shapes. In: Proc. IEEE Comput.Soc. Conf. Computer Vision and Pattern Recognition, vol. 1, pp. 723–730 (2001)
18. Taylor, C.J.: Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In: CVIU, vol. 80, pp. 349–363 (2000)
19. Remondino, F., Roditakis, A.: 3D Reconstruction of Human Skeleton from Single Images or Monocular Video Sequences. In: Michaelis, B., Krell, G. (eds.) DAGM 2003. LNCS, vol. 2781, pp. 100–107. Springer, Heidelberg (2003)
20. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: IJCAI, pp. 674–679 (1981)

Personal Recognition Using Single-Sensor Multimodal Hand Biometrics

Andreas Uhl and Peter Wild

University of Salzburg, Department of Computer Sciences, A-5020 Salzburg, Austria
{uhl,pwild}@cosy.sbg.ac.at

Abstract. Single-sensor approaches to multimodal biometric authentication targeting the human hand in multiple-matcher scenarios provide higher security in terms of accuracy and resistance to biometric system attacks than unimodal systems. This paper introduces a novel multimodal hand biometric system using palmar images acquired by a commercially available flatbed scanner. Hence, the presented approach to personal recognition is independent of specific biometric sensors, such as fingerprint readers or palmprint scanners. Experimental results with a minimum half total error rate of 0.003% using a database of 443 hand images will illustrate the performance improvement when hand-geometry, fingerprint and palmprint-based features are combined.

1 Introduction

Biometrics facilitates authentication tasks by means of independence of knowledge and tokens (which may be passed on fraudulently), permanent and singular features (which may hardly be lost without intention nor forgotten like passwords), and low mean transaction times for access control. Biometric authentication systems in verification mode (comparison with a claimed identity) should not exceed 0.1% False Match Rate (FMR) at 0.1% False Non-Match Rate (FNMR) according to [1]. In order to bridge the gap between required rates and performance of current unimodal matchers, multimodal systems combine different modalities. The contribution of this paper is a novel single-sensor multibiometric system integrating common modalities related to the human hand:

- *Hand geometry* using a feature targeting widths of single fingers;
- *Fingerprint* extracting local-level textural features at fingertips;
- *Palmprint* focusing on textural properties of the human palm tracking principal lines and wrinkles;

The novelty of the presented approach lies in the extraction of all features out of a simple palmar scan using a common flatbed scanner, which will be shown to produce very accurate results. Originally different sensing devices have been developed for each of the integrated modalities. The presented approach is motivated by Kumar et al. [2]. However, their results only refer to randomly paired samples from multiple databases indicating an improvement of the best single

biometric trait by means of Equal Error Rate (EER) performance of roughly 40% (yielding a total EER of 3.53%). Recently, an implemented solution extracting samples out of a single acquired input signal based on multispectral image processing has been presented by Rowe et al. [3]. They report no errors on their dataset of 50 volunteers for fingerprint score fusion using the Sum Rule [4] and a significant performance improvement if ring finger and palmprint scores are combined. A significant drawback of their presented solution is the requirement of special image sensors based on multispectral technology. Using document scanners designed for large markets instead has several advantages:

- *Availability*: according to a study in [5], every third US household held a document scanner by 2001 expecting a market saturation by 2006.
- *Reproducibility*: sensor requirements of 500-dpi 8-bit grayscale at reasonable scanning speed are largely covered by existing scanners.
- *Sensor independence*: results of the employed HP Scanjet 3500c model, which is a Charge Coupled Device (CCD) scanner placed 2002 in the low-cost workgroup market segment [5], are most likely to remain stable or even improve when faster and more accurate hardware is employed.
- *Cost*: when hardware for large markets is used, acquisition and upkeep costs are minimized.

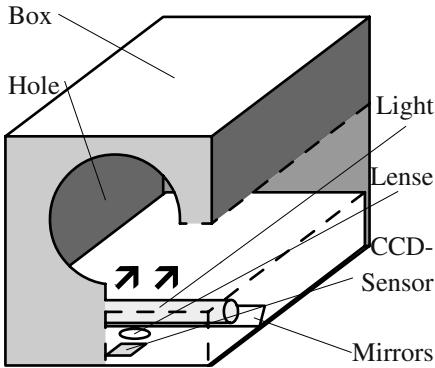
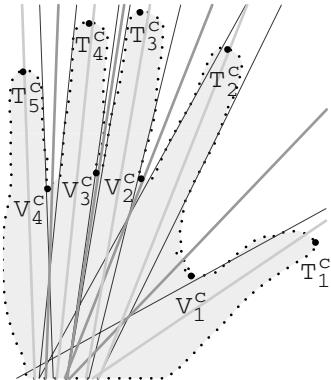
The structure of this paper follows the information flow in common biometric systems. After a description of sensor and preprocessing steps in Sect. 2, feature extraction and matching techniques are introduced in Sect. 3. System performance is assessed by experimental evaluation in Sect. 4. Results are discussed comparing single and combined modalities. Section 5 forms the conclusion.

2 Sensing and Preprocessing

For experiments we used a HP Scanjet 3500c scanner in 500-dpi and 8-bit grayscale mode. Users were free to choose an arbitrary placement when inserting their hand through a hole in a box containing the sensor (to minimize environmental light, see Fig. 1) as long as their fingers did not touch each other, i.e. the system is *peg-free*. Thus, preprocessing is essential to provide each extractor with normalized data and to achieve rotational and translational invariance.

After input acquisition, the hand object is localized within the 4250×5850 image with respect to the smallest circumscribed axis-aligned rectangle. The $a \times b$ hand image $H(x, y)$ with $1 \leq x \leq a, 1 \leq y \leq b$ is segmented using a fast version of Otsu's thresholding [6]. Arm parts in the input image are removed iteratively. Within every step of the iteration, the hand image is aligned using statistical moments estimating the inclination of the best-fitting ellipse centered in the center of mass, a technique from face-detection [7]. Then, the horizontal wrist line separating arm and palm is estimated by a top-down scan and the corresponding arm part is masked in the input image. This procedure is iterated 3 times to obtain a stable segmentation.

Despite its application in many hand biometric systems, such as in [8, 9], ellipse-fitting is found to produce imperfect alignment for different spreadings of

**Fig. 1.** Employed flatbed sensor**Fig. 2.** Salient point refinement for hands

ingers. Instead, we employ a predefined hand coordinate system, an alignment procedure frequently used in palmprint recognition [10]. Let \mathbf{V}_i with $1 \leq i \leq 4$ denote the finger valleys of a right hand in counter-clockwise order and let \mathbf{T}_j with $1 \leq j \leq 5$ track positions of finger tips of thumb, index, middle, ring and small finger. Then the origin of the new coordinate system is defined as the position of the inter-finger valley \mathbf{V}_3 and its y-axis is the line parallel to the least-squares approximation of the outer palm boundary through \mathbf{V}_3 . The palm boundary is defined as the subsequence $\langle \mathbf{C}_s, \dots, \mathbf{C}_e \rangle$ of the counter-clockwise contour sequence $\langle \mathbf{C}_0, \dots, \mathbf{C}_z \rangle$ within a range defined by the average length l of fingers for the given hand:

$$s = \max\{k \in \{0, \dots, z\} : \|\mathbf{C}_k - \mathbf{T}_5\| \leq \|\mathbf{V}_4 - \mathbf{T}_5\|\}; \quad (1)$$

$$e = \max\{k \in \{0, \dots, z\} : \|\mathbf{C}_k - \mathbf{C}_s\| \leq l\}. \quad (2)$$

Candidates for salient points (i.e. finger peaks and valleys) are typically extracted from the silhouette contour by finding local extrema of the radial distance function [11] with respect to a reference point at the wrist. In our approach, peak candidates are refined by intersecting the major axis of the best-fitting ellipse for each individual finger (using candidate valleys as separators between fingers) with the contour. Finger valley refinement comprises an approximation of the proximal finger boundary by straight lines on each side and an intersection of bisecting lines between fingers with the contour, see Fig. 2.

3 Feature Extraction and Matching

We employ four different geometric and texture-based algorithms:

- *Shape*: divides each finger into slices of equal height and keeps track of the local finger width for each slice (see [12] for an application on footprints).

- *Minutiae*: identifies minutiae points within a finger-axis-aligned rectangle at finger tips using NIST’s minutiae extraction software [13].
- *Palmprint*: extracts variances of 144 overlapping blocks in the palmprint image after edge detection (method introduced by Kumar et al. [9]).
- *Eigenpalms and Eigenfingers*: projects each finger and palm onto the eigen-space spanned by the most significant principal components (see [14]).

The employed scenario represents a score level fusion approach (i.e. matching scores rather than feature vectors or decision vectors of independent algorithms are combined) operating in multiple-matcher configuration. Whereas more independent levels of integration exist (such as multiple biometrics or multiple sensors for example), information can be expected to be largely independent, since employed algorithms operate at:

- *different resolutions*: Shape, Minutiae and Palmprint require the full 500-dpi input signal, Eigenpalms and Eigenfingers are extracted at 100 dpi.
- *different parts of the hand*: Minutiae operates on each of the five finger tips, Eigenpalms and Palmprint extract information from the textural palm area while Eigenfingers and Shape concentrate on fingers.

Each individual feature extractor contributes its extracted representation to a common augmented feature vector. This template is stored together with an identifying key in a preceding enrollment mode and compared with the stored reference in authentication mode [4]. For matching, augmented feature vectors are decomposed into feature-dependent parts and matching scores are generated for each of the employed features.

3.1 Shape Feature

The Shape feature depicted in Fig. 3 targets hand geometry measurements extracting local finger widths from the hand contour. Each finger is rotated upright

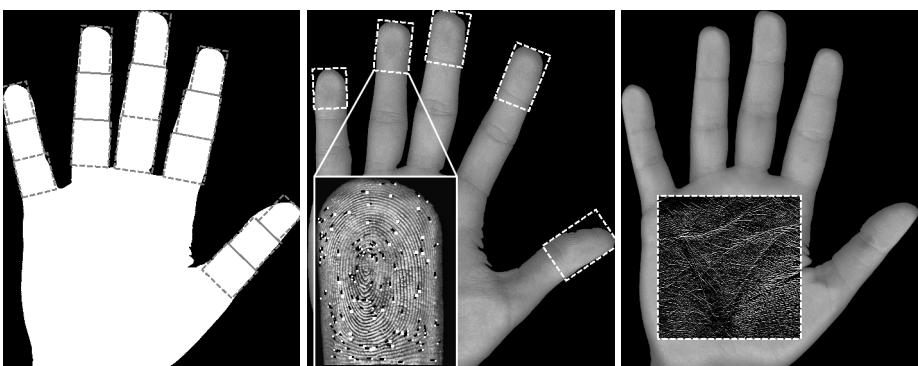


Fig. 3. Shape feature

Fig. 4. Minutiae feature

Fig. 5. Palmprint feature

using moments [7] and cropped at the adjacent finger valley closer to the tip. For each of $c = 3$ slices S_1, \dots, S_c of equal height covering the finger we determine the object's average width $w(S_n)$, $1 \leq n \leq c$, with respect to the y-monotone contour extracted using a left-right scan. Matching is based on Manhattan metric.

3.2 Minutiae

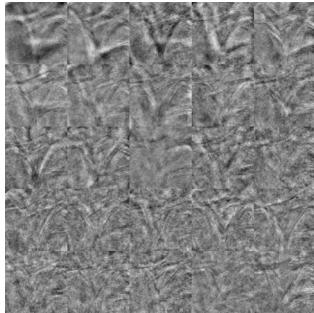
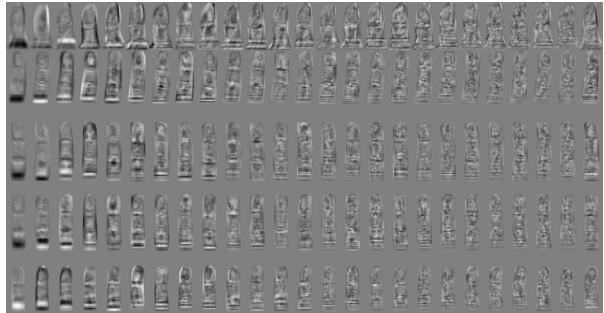
We extract position, orientation and quality of termination and bifurcation points in the outermost structural part of the epidermis employing the NFIS2 [13] extraction software *mindtct* on the enhanced fingerprint image. While sensing regions may also be defined for other parts of the human hand (see [3] for results on metacarpal skin), we extract each fingerprint image as a $w \times \frac{h}{3}$ (and $w \times \frac{h}{2}$ for the thumb, respectively) rectangular area aligned with respect to the major axis of the finger circumscribed by its $w \times h$ sized bounding box. Since flatbed optical images exhibit low contrast between ridges and valleys, Contrast-limited adaptive histogram equalization described in [15] has been employed. Minutiae pairing and matching is executed using the NFIS2 *bozorth3* matcher. Matching results of individual fingerprints are combined using the Max Rule (returning the highest score of all fingers) [4].

3.3 Palmprint Feature

After localization of the hand-axis aligned square palmprint region of size l (being the average finger length) centered in $\mathbf{P} = (0, -\frac{7}{10}l)$ with respect to the introduced hand coordinate system, edges are detected using a 7×7 -Prewitt filter, see Fig. 5. Features are extracted as variances within blocks of size 24×24 pixels within a downsampled and normalized 300×300 version of the palm, as proposed in [9]. The Palmprint matcher computes a normalized squared Euclidian distance between feature vectors.

3.4 Eigenpalms and Eigenfingers

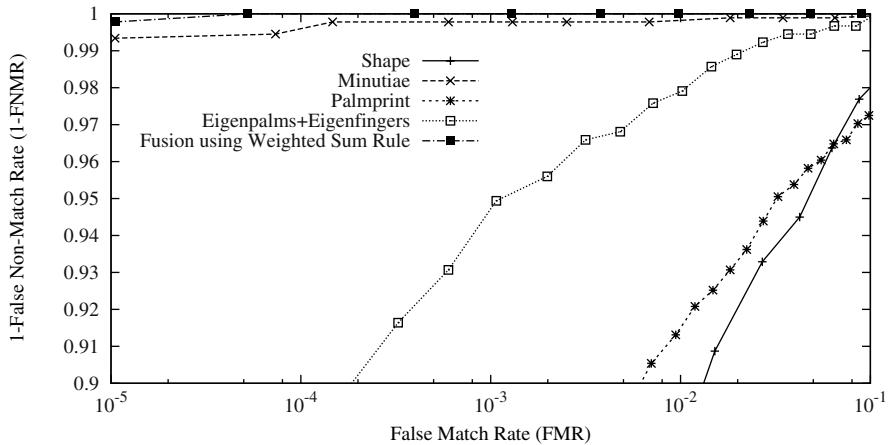
Eigenpalms and Eigenfingers (Figs. 6, 7) are derived from Eigenfaces introduced by Turk and Pentland [16] and are based on the Karhunen-Loeve transform converting image vectors of single fingers or palms into a space spanned by the most significant components calculated from a set of training images. After subtracting the average image vector \mathbf{A} for each type of training image (256×256 palm images, 128×384 index, middle, ring fingers and 128×256 thumb and little finger images), the $d = 25$ eigenvectors \mathbf{E}_m , $1 \leq m \leq d$ with largest eigenvalues are computed from the covariance matrix of normalized samples. For each image type feature extraction corresponds to (a) extraction of finger or palm image \mathbf{I} ; (b) normalization with respect to the average image vector \mathbf{A} and; (c) projection onto the corresponding space spanned by \mathbf{E}_m and storage of projection coefficients as feature components. The corresponding matcher is based on Manhattan metric, matching scores of individual eigenspaces are combined using the Product Rule.

**Fig. 6.** Eigenpalms**Fig. 7.** Eigenfingers

4 Performance Evaluation

We evaluate verification performance of single and combined features in the introduced hand biometric system by means of Receiver operating characteristics depicted in Fig. 8 and indicators given in Table II including MinHTER performance, i.e. the global minimum of the *Half total error rate* function [17]:

$$\text{HTER}(t) := \frac{\text{FMR}(t) + \text{FNMR}(t)}{2}. \quad (3)$$

**Fig. 8.** Receiver operating characteristics

4.1 Experimental Database

For testing purposes a database of 443 right-hand samples of 71 males and 15 females recorded at the Department of Computer Sciences, University of

Table 1. Verification results for the employed algorithms

Algorithm	MinHTER	ZeroFMR	ZeroFNMR
Shape	4.7%	70.74%	25.53%
Minutiae	0.12%	1.1%	16.44%
Palmpoint	4.1%	36.19%	100%
Eigenpalms + Eigenfingers	1.44%	15.29%	10.72%
Fusion using Weighted Sum Rule	0.003%	0.33%	0.005%

Salzburg, Austria has been employed. Samples were acquired within a time-span of 15 minutes per person. All users were allowed to wear rings or watches. Failure to Acquire (FTA) exceptions occurred in 0.9% of the cases.

4.2 Verification Performance

When executing a cross-comparison of available templates resulting in 909 genuine and 95232 imposter comparisons, the best matching results with respect to MinHTER performance are returned by the Minutiae feature with 0.12%. Especially for high-security environments, where low FNMR at almost zero FMR is demanded, this feature outperforms all other tested algorithms. Rates for fusion-based Eigenpalms + Eigenfingers-based recognition are an order of magnitude higher with a MinHTER value of 1.44%. Low resolution requirements of 100 dpi, fast matching speed and reasonable matching performance especially for applications demanding high convenience are special characteristics of this feature. However, recently published results by Cheung et al. [18] indicate that recognition rates for Eigenpalms tend to degrade significantly in case of larger time lapses between recordings and twins. The Palmpoint feature provides less accurate results at 4.1% MinHTER, which lies in the order of reported error rates in [9]. Geometry-based Shape could not outperform texture-based features with 4.7% MinHTER. If results from all modalities are combined using Weighted Sum Rule fusion, a significantly higher performance of 0.003% MinHTER can be achieved. An optimum configuration of weights (with respect to a step size of 0.01) was found at 0.10 for Shape, 0.17 for Palmpoint, 0.06 for Eigenpalms + Eigenfingers, and 0.67 for Minutiae. Although the eigenspace-based features exhibit the second best MinHTER-performance, their weight within the best common feature is rather small. This behavior might possibly be caused by high correlation between Eigenfingers and Shape, since Eigenfingers are sensitive to both shape and textural information.

5 Summary

We have proposed a new multimodal flatbed-scanner-based approach to personal recognition with hands integrating hand geometry, fingerprint and palmpoint-based algorithms in a multiple-matcher scenario. The performance of the introduced system has been assessed using a custom medium-sized database.

Considering the overall performance in verification mode a MinHTER of 0.003% could be achieved with fusion at matching score level, outperforming each of the individual features. Generally, texture-based features produced better results (Minutiae with 0.12% MinHTER, followed by Eigenpalms + Eigenfingers with 1.44% and Palmprint with 4.1%) than geometry-based Shape with 4.7% MinHTER. Future topics of interest comprise larger time-lapses between recordings, studies on twins and an evaluation based on more different flatbed sensors.

References

1. Jain, A.K., Pankanti, S., Prabhakar, S., Hong, L., Ross, A.: Biometrics: A grand challenge. In: Proceedings of the 17th International Conference on Pattern Recognition, pp. 935–942 (2004)
2. Kumar, A., Zhang, D.: Combining fingerprint, palmprint and hand-shape for user authentication. In: Proceedings of the 18th International Conference on Pattern Recognition, pp. 549–552 (2006)
3. Rowe, R.K., Uludag, U., Demirkus, M., Parthasaradhi, S., Jain, A.K.: A multispectral whole-hand biometric authentication system. In: Proceedings of Biometrics Symposium, pp. 1–6 (2007)
4. Ross, A.A., Jain, A.K.: Information fusion in biometrics. Pattern Recognition Letters 24, 2115–2125 (2003)
5. InfoTrends, Inc.: Scanner market reaches maturity - penetration nearing one third of U.S. PC households (2001), <http://www.infotrends.com/public/Content/Press/2001/06.19.2001.html>
6. Liao, P.S., Chen, T.S., Chung, P.C.: A fast algorithm for multilevel thresholding. Journal of Information Science and Engineering 17(5), 713–727 (2001)
7. Sobottka, K., Pitas, I.: Extraction of facial regions and features using color and shape information. In: Proceedings of the 13th International Conference on Pattern Recognition, pp. 421–425 (1996)
8. Yörük, E., Dutagaci, H., Sankur, B.: Hand biometrics. Image and Vision Computing 24(5), 483–497 (2006)
9. Kumar, A., Wong, D.C.M., Shen, H.C., Jain, A.K.: Personal verification using palmprint and hand geometry biometric. In: Proceedings of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication, pp. 668–678 (2003)
10. Zhang, D.: Palmprint authentication. Kluwer Academic Publishers, Dordrecht (2004)
11. Yörük, E., Konukoglu, E., Sankur, B., Darbon, J.: Shape-based hand recognition. IEEE Transactions on Image Processing 15, 1803–1815 (2006)
12. Uhl, A., Wild, P.: Personal identification using eigenfeet, ballprint and foot geometry biometrics. In: Proceedings of the IEEE First International Conference on Biometrics: Theory, Applications, and Systems, pp. 1–6 (2007)
13. NIST: Fingerprint Image Software 2 (2004), <http://fingerprint.nist.gov/NFIS>
14. Ribaric, S., Fratric, I.: A biometric identification system based on eigenpalm and eigenfinger features. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(11), 1698–1709 (2005)
15. Zuiderveld, K.: Contrast limited adaptive histogram equalization. In: Graphics gems IV, San Diego, CA, USA, pp. 474–485. Academic Press Professional, Inc, London (1994)

16. Turk, M., Pentland, A.: Eigenfaces for Recognition. *Journal of Cognitive Neuroscience* 3(1), 71–86 (1991)
17. Pierrot, J., Lindberg, J., Koolwaaij, J., Hutter, H.P., Genoud, D., Blomberg, M., Bimbot, F.: A comparison of a priori threshold setting procedures for speaker verification in the CAVE project. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 125–128 (1998)
18. Cheung, K.H., Kong, A., Zhang, D., Kamel, M., You, J.: Does eigenpalm work? a system and evaluation perspective. In: *Proceedings of the 18th International Conference on Pattern Recognition*, pp. 445–448 (2006)

Face Tracking for Spatially Aware Mobile User Interfaces

Jari Hannuksela, Pekka Sangi, Markus Turtinen, and Janne Heikkilä

Machine Vision Group, Infotech Oulu

Department of Electrical and Information Engineering
P.O. Box 4500, FIN-90014 University of Oulu, Finland
jari.hannuksela@ee.oulu.fi

Abstract. This paper introduces a new face tracking approach for controlling user interfaces in hand-held mobile devices. The proposed method detects the face and the eyes of the user by employing a method based on local texture features and boosting. An extended Kalman filter combines local motion features extracted from the face region and the detected eye positions to estimate the 3-D position and orientation of the camera with respect to the face. The camera position is used as an input for the spatially aware user interface. Experimental results on real image sequences captured with a camera-equipped mobile phone validate the feasibility of the method.

Keywords: facial feature extraction, motion analysis, pose estimation.

1 Introduction

Modern mobile communication devices are attractive platforms for various new applications as their multimedia capabilities are improving together with their computational resources. As more and more applications are available for these devices, the user interfaces are becoming overloaded, potentially confusing the user who needs to learn to use each individual application. Particularly, the small displays set the primary restrictions for the usability. A viable alternative to improve user interaction with mobile terminals are spatially aware displays [1]. The solution is to provide a window on a larger virtual workspace where the user can access more information by moving the device around.

Based on this approach, Yee [2] presented a peephole display technique to control, for example, the 3-D image viewer and the 3-D calendar applications on a hand-held computer using several additional sensors providing positional information. Recent work on small sized devices such as mobile phones has focused on employing motion data measured using the accelerometer [3] or the camera [4] for this same purpose. The motion can be integrated in order to obtain a positional input for controlling the spatially aware display. However, when integrating motion data, errors typically accumulate over time. Therefore, motion based techniques should be used only when a precise position is not mandatory. Our idea is to use a built-in camera as a sensor to determine the position and orientation (pose) of the device with respect to the user face and utilize this information for browsing the virtual workspace of the mobile terminal in its display. The pose is measured continuously, which means that the user interface is spatially aware of the user all the time.

A number of face and head pose estimation methods have been proposed mainly for personal computers. Prior research related to our method include early work of Azarbayejani et al. [5]. They used recursive estimation to recover a head structure and motion from image sequences of rigid motion. The extended Kalman filter (EKF) was applied to track distinct features corresponding to the corners of the eyes and nostrils. Black and Yacoob [6] developed a regularized optical flow method that used an eight parameter 2-D planar model. Based on their work, Basu et al. [7] presented a system to track heads with a large amount of head motion. Instead of using a planar model they used a 3-D ellipsoidal model of the head. The algorithm starts by computing optical flow for the image and then it estimates the 3-D motion of the rigid model. The system was not real-time due to slow computation of the optical flow. La Cascia et al. [8] presented an approach using a texture mapped 3-D rigid surface model for the head and formulated tracking as an image registration problem.

In this paper, we propose a method for 3-D face tracking that can be used to control spatially aware user interfaces of mobile devices. Unlike many other methods proposed in the literature, the low computational cost of our method makes it practical for mobile platforms where high computational resources are not available. Also, we wish to emphasise the point here that our applications differs from the usual case since the device is moved with respect to the face. The proposed system consists of two stages. In the initialization stage, the users face and eyes are detected automatically. Whenever a face is detected the tracking is started. During tracking, an extended Kalman filter estimates the camera pose utilizing a novel combination of motion features and eye positions detected from the face region.

2 Facial Feature Extraction

The automatic facial feature extraction starts with a face detection process. After the face of the user is found, the eyes are searched for within the face region. We use the detected eye positions to initialise the face model for the tracker. In the tracking phase, we detect eyes using the same method and also extract motion features which provide information about local motion for different parts of the facial region.

2.1 Face Detection

Our object detection approach uses efficient gray-scale invariant texture features [9] and boosting [10]. The solution is based on the local binary pattern (LBP) texture methodology [9] for facial representation. The AdaBoost learning is then applied to find the most discriminative features for distinguishing the face patterns from the background.

The LBP features can detect local texture primitives such as spots, edges or corners from the images [9]. They have been found to be very discriminative in facial image analysis [11]. Similar to the approach of Hadid et al. [12], we use the facial representation where the LBP feature histograms are separately computed over the sparse set of local image regions and the whole face area. These are then concatenated for creating the final face descriptor. Fig. 1 illustrates the idea of LBP based facial representation.

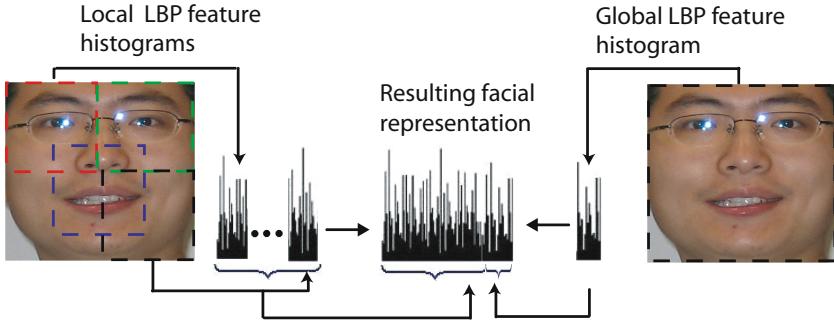


Fig. 1. Facial representation with LBP texture features: a face image is with set of local and global LBP histograms

The original 20×20 pixels face image is divided maximally into nine overlapping regions of 10×10 pixels. The basic 4-neighbors LBP operator with 16-bin histogram is used resulting totally of 144-bin histogram of local features. In addition to that the global 16-bin histogram is concatenated to the local features in order to create the full 160-bin face histogram.

The AdaBoost algorithm is a discriminative learning method that has been widely used in different object detection tasks. The idea is to combine several relatively weak features into one stronger hypothesis [10]. In our case the weak features are the LBP histogram bin values calculated over certain image regions. These positions and corresponding LBP values are learned off-line with the set of labeled face samples. Few example training images are shown in Fig. 2a.

Inspired by the well known Viola and Jones [13] object detection approach, we built a cascaded classifier structure to speed up the detection. In the early cascade levels only a few histogram positions were considered to rapidly reject majority of classified image regions. The face like image regions were classified with more features in order to make robust detection. We applied the trained cascade to the image pyramid using a sliding window approach to classify image regions in different scales. As a result, the rectangular coordinates of each detected face was obtained. Example output from the face detection algorithm is shown in Fig. 3.

2.2 Eye Detection

Once a possible face of the user is detected, the eyes are searched for within the facial region. The approach for detecting the eyes is similar to the face detection algorithm except the size of the template image is 16×20 pixels to more accurately cover the eye region. Some example training samples are shown in Fig. 2b. Face orientation and size were used to restrict the search window and scales in the eye detection phase. In this sense, the actual detection is carried out with very low computational costs. The centers of the detected windows are used as eye coordinates for the tracker.

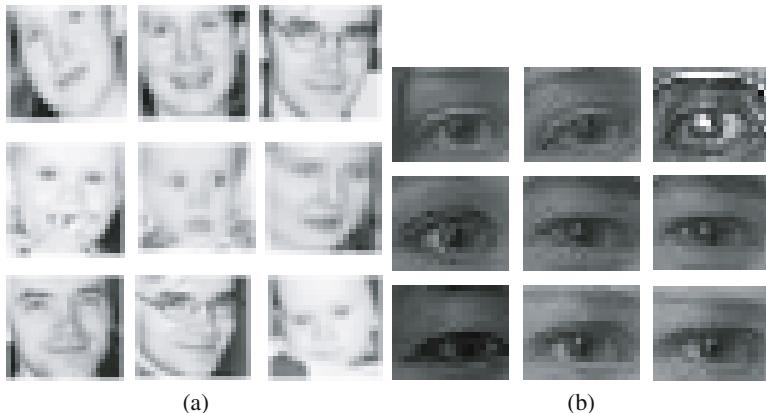


Fig. 2. Example training images for face (a) and eye (b) detector

2.3 Motion Features

In motion feature extraction, we apply an efficient method developed in our earlier work [14]. First, the face region of the previous frame is split to 16 subregions, and one 8x8 pixel block is selected from each region based on analysis of spatial gradient information. Displacement of selected blocks is estimated using zero mean sum of squared differences (ZSSD) criterion which is applied over some range of candidate displacements (e.g. 16 pixels). Refinement to subpixel precision is done in the neighborhood of the displacement minimizing the ZSSD measure using fitting of the second order polynomials.

The ZSSD values are also used to analyse uncertainty information related to the local displacement estimate. This information is used in RANSAC style outlier analysis which provides reliable motion features for 3-D pose estimation. Each motion feature consists of (1) block position in the previous frame \mathbf{p}_i (2) block displacement estimate \mathbf{d}_i and (3) displacement error estimate encoded as a 2×2 covariance matrix \mathbf{C}_i . This information in addition to the eye coordinates is used as measurements in the tracking phase.

3 Face Tracking and Pose Estimation

The 3-D camera pose with respect to the face can be estimated based on a face model and corresponding 2-D image observations. In our method, the eyes and motion features detected from the face region are used as measurements. We apply extended Kalman filtering (EKF) to estimate the camera position and to track the eyes as well as the motion features.

3.1 Face Pose and Motion Model

We model the face as a rigid plane which is not an accurate description for the face, but considering our application it provides sufficient pose estimates to control the user

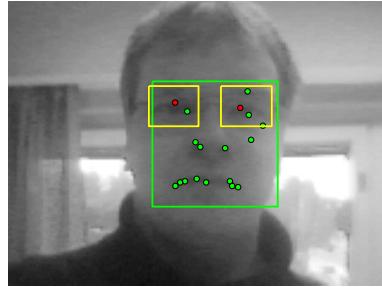


Fig. 3. Facial features: face detection result (green rectangle), eye detection result (yellow rectangles and red dots), centers for the eyes (red dots) and motion features (green dots)

interface. The center of the face coordinate system is set to the center of the detected face region in the first frame. Initially the face plane is parallel to the image plane and its distance to the camera is some predefined constant. The face model includes the eye positions (\mathbf{P}_1 and \mathbf{P}_2) and the motion feature positions ($\mathbf{P}_i, i = 3, 4, \dots, N + 2$). \mathbf{P}_1 and \mathbf{P}_2 are the backprojected face plane coordinates of the eyes detected in the first image. $\mathbf{P}_i, i = 3, 4, \dots, N + 2$ are the backprojected face plane coordinates of the motion feature positions \mathbf{p}_i where projection uses the pose estimate associated with the previous frame. The model points for motion features are updated for each incoming frame, but the eye positions are fixed after initialization.

The first order dynamic model for the system is represented as

$$\mathbf{x}_{k+1} = \Phi_k \mathbf{x}_k + \Gamma_k \varepsilon_k, \quad (1)$$

where the pose and the velocity terms are included in the state vector \mathbf{x}_k at time instant k , and Φ_k is the corresponding state transition matrix. Specifically, the state vector \mathbf{x}_k consists of the position (x_k, y_k, z_k) and orientation $(\omega_k, \varphi_k, \kappa_k)$ of the face model with respect to the camera and the corresponding velocities. It is defined as follows

$$\mathbf{x}_k = [x_k, \dot{x}_k, y_k, \dot{y}_k, z_k, \dot{z}_k, \omega_k, \dot{\omega}_k, \varphi_k, \dot{\varphi}_k, \kappa_k, \dot{\kappa}_k]^T.$$

$\Gamma_k \varepsilon_k$ models the uncertainty of the dynamic model. Γ_k is the process noise transition matrix. We model arbitrary accelerations as process noise ε_k which is assumed to be Gaussian distributed with zero-mean and the covariance matrix $\mathbf{Q}_k = E\{\varepsilon_k \varepsilon_k^T\}$. The pose parameters are initialized according to the initial face model. In the beginning, the velocities are set to zero. We approximate the variances of the process noise from the maximum accelerations assumed.

3.2 Measurements

The measurement model is needed to relate the 3-D pose parameters (state) to the 2-D image observations. We use a perspective camera model to transform the object coordinates \mathbf{P}_i to image coordinates (u_i, v_i) and the model is

$$\mathbf{z}_k = \mathbf{h}(\mathbf{P}_i, \mathbf{x}_k) + \eta_k, \quad (2)$$

where $\mathbf{h}(\mathbf{P}_i, \mathbf{x}_k)$ is a non-linear observation function that uses the perspective camera model to transform object coordinates \mathbf{P}_i to image coordinates \mathbf{z}_k using state \mathbf{x}_k . The observation model is linearized using its partial derivatives with respect to the state variables \mathbf{x}_k to obtain the Jacobian matrix.

The actual measurement $\mathbf{z}_i, i = 1, \dots, 2$ for the eyes are extracted in a region around the predicted locations. The size for the region is adjusted dynamically by projecting the estimation uncertainty to the image coordinates. The eye detection is performed as described in Sec. 2.2 and (u, v) -coordinates are obtained as a result. If there is too much deviation from the prediction, then the predicted position is chosen instead of the actual measurement. In the case of motion features, the measurement is $\mathbf{z}_i = \mathbf{p}_i + \mathbf{d}_i, i = 3, \dots, N + 2$, where \mathbf{p}_i is motion feature position in the previous frame and \mathbf{d}_i is the detected displacement in the current frame. The measurement noise η_k for the eyes is assumed to be Gaussian with zero-mean and variance of 9 pixels. The error covariance matrix \mathbf{C}_i is derived for each feature separately using the method described in [14].

3.3 Tracker

The EKF algorithm estimates the pose recursively repeating prediction and correction stages. In the first stage, the pose and locations of the features at the next time instant are predicted based on the previous pose estimate and the dynamical model. In the correction stage the predicted pose is adjusted by using facial feature measurements. Also, the face region is updated based on the estimated pose allowing motion feature extraction from the correct image region. The measurements can occasionally cause errors that make the tracking fail. If the feature position is not within the uncertainty limits derived from the error covariance matrix or the feature is lost then the prediction is used instead of the measurement. The error cannot exist longer than three frames or the initialization is started again.

4 Experiments

We evaluated the feasibility of the method for controlling the spatially aware user interface with a camera-equipped mobile phone. The phone used was Nokia N95 containing ARM11 based processor running at 330 MHz. It has two built-in cameras, one for high resolution imaging and other low resolution camera pointing to the users' face. In experiments, we used the latter with an image resolution of 320x240 pixels and a maximum obtainable frame rate of 15 fps.

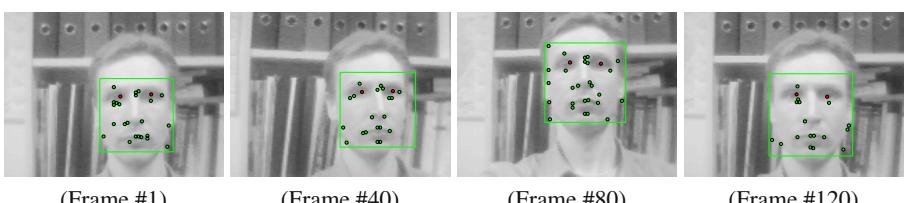


Fig. 4. Facial feature tracking example

In the experiments, we asked a subject to perform rectangular hand movement representing hypothetical input for the user interface. In order to test repeatability of the method, a rectangle (120x90 mm) was first drawn on the board. The user followed the outline of the rectangle and repeated the experiment 10 times. The trajectory for the movement was obtained by solving the camera position from the estimated 3-D pose of the face. Fig. 5(a) shows the obtained 3-D trajectories of the device as a projection in the $x - y$ plane and Fig. 5(b) illustrates the same result in 3-D coordinates. The trajectories are uniformly scaled and aligned with each other. As illustrated in the figure, the repeatability achieved for the method is reasonably good and it allows the position based control of user interfaces. The results can be further improved by adding more fixed points other than the eyes to the face model providing enhanced positional information. Possible additional points are, for example, the mouth, the mouth corners and the nostrils.

We also assessed the feature tracking subjectively because the ground truth for the pose was not known. Fig. 4 shows an example result of the extracted facial feature locations during tracking. The tracking was successfully completed without interruptions in all of the sequences. There were sometimes large errors in the eye measurements or rarely the eye was not detected at all. In these situations the tracker used the predicted measurements and tracking did not fail. Although the lighting changed continuously throughout the sequences, the system still worked properly.

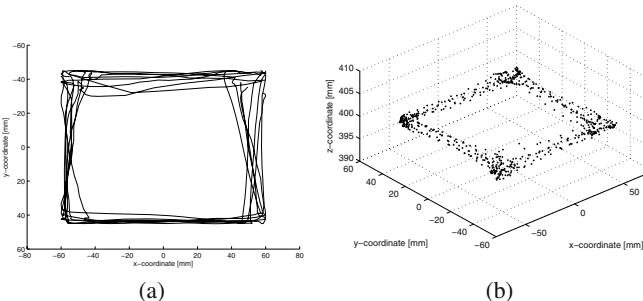


Fig. 5. Trajectories for the repeatability test: (a) trajectory in x - y plane (b) trajectory in 3-D coordinates

5 Conclusions

We have presented a new feature based head tracking method for controlling user interfaces of mobile devices. It automatically detects the face and eyes of the user employing local texture features and boosting. The 3-D position of the device with respect to the user face is estimated with extended Kalman filtering using a novel combination of motion features and eyes extracted from the face region. The camera position obtained can be used to control the spatially aware user interface of camera-equipped mobile phone. The advantage of the method is that real-time performance on a resource limited mobile device can be easily achieved, which makes it useful for practical applications.

Acknowledgments

The financial support of the Academy of Finland (project no. 110751) is gratefully acknowledged.

References

1. Fitzmaurice, G.W.: Situated information spaces and spatially aware palmtop computers. *Communications of the ACM* 36(7), 38–49 (1993)
2. Yee, K.P.: Peephole displays: pen interaction on spatially aware handheld computers. In: Proc. of the SIGCHI conference on human factors in computing systems, pp. 1–8 (2003)
3. Hinckley, K., Pierce, J., Sinclair, M., Horvitz, E.: Sensing techniques for mobile interaction. In: Proc. 13th ACM symposium on User Interface Software and Technology, pp. 91–100 (2000)
4. Haro, A., Mori, K., Capin, T., Wilkinson, S.: Mobile camera-based user interaction. In: Proc. of IEEE Workshop on Human-Computer Interaction, pp. 79–89 (2005)
5. Azarbayejani, A., Starner, T., Horowitz, B., Pentland, A.: Visually controlled graphics. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15(6), 602–605 (1993)
6. Black, M.J., Yacoob, Y.: Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In: Proc. IEEE International Conference on Computer Vision, pp. 374–381 (1995)
7. Basu, S., Essa, I., Pentland, A.: Motion regularization for model-based head tracking. In: Proceedings of the 13th IEEE International Conf. on Pattern Recognition, pp. 611–616 (1996)
8. La Cascia, M., Sclaroff, S., Athitsos, V.: Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 22(4), 322–336 (2000)
9. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(7), 971–987 (2002)
10. Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Europ. Conf. on Computational Learning Theory, pp. 23–37 (1995)
11. Hadid, A., Zhao, G., Ahonen, T., Pietikäinen, M.: Face analysis using local binary patterns. In: Mirmehdi, M. (ed.) *Handbook of Texture Analysis*. Imperial College Press (2007)
12. Hadid, A., Pietikäinen, M., Ahonen, T.: A discriminative feature space for detecting and recognizing faces. In: IEEE Conference on Computer Vision and Pattern Recognition (2004)
13. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE International Conf. on Computer Vision and Pattern Recognition, pp. 511–518 (2001)
14. Sangi, P., Hannuksela, J., Heikkilä, J.: Global motion estimation using block matching with uncertainty analysis. In: 15th European Signal Processing Conference (2007)

Markerless Augmented Reality Using Image Mosaics

Pietro Azzari, Luigi Di Stefano, Federico Tombari, and Stefano Mattoccia

ARCES - DEIS, University of Bologna,

viale Risorgimento 2, 40125 Bologna, Italy

{pazzari,ldistefano,ftombari,smattoccia}@deis.unibo.it

<http://www.vision.deis.unibo.it/>

Abstract. Augmented reality is a powerful tool for delivering spatially coherent information to a user moving in a known environment. Accurate and reliable pose estimation is the key to success. Many approaches track reference objects into the scene but as the environment grows larger more objects need to be tracked leading to computationally intensive methods. Instead, we propose a practical approach that is suitable for environment where big planar structures are present. All the objects laying on the structure are composed into a large reference object using image mosaicing techniques, so that the problem is reduced to that of finding the pose from a single plane. Experimental results show the effectiveness of this approach on two interesting case studies such as aeronautical servicing and cultural heritage.

Keywords: Mosaicing, Augmented reality, Pose estimation, Markerless.

1 Introduction

Augmented reality techniques convey information that is both semantically and spatially coherent with the observed scene. Information is shown by augmenting the scene captured through a camera with graphical objects that are properly aligned with the world 3D structure and often contextually close to the user needs. In this paper we mainly focus on structural coherence, nonetheless a simple demonstration of contextual awareness is given in the experimental results section.

The capability to deliver spatially coherent information to a user moving in a known environment is enabled by accurate and reliable pose estimation algorithms. Such algorithms try to compute the pose of the observer with respect to the world he is moving in by establishing correspondences among objects detected in the scene. Using those correspondences both the information to be displayed and the structure of the scene is estimated.

Most of the algorithms described in literature can be thought of in terms of a binary taxonomy: those that rely on absolute information [12], such as known models, and those based on chained transformations [34]. The former seek to find camera poses that correctly reproject some fixed features of a given 3D

model into the $2D$ image. They do not drift but they often lack precision which results in jitter. The latter do not exploit a priori information but match interest points between images. Since the correspondences between adjacent frames are usually located very precisely, these algorithms do not jitter but suffer from drift or even loss of track.

Pose estimation algorithms usually represent the world as a collection of reference objects, modeled as $3D$ meshes, associated with appearance models, such as collection of key frames or image patches related to each vertex. Navigation of large environments is handled using several objects so that many of them are visible even though the user moves widely across the environment. Many algorithms are known to estimate the pose very quickly using a single object and a single image [12]. However, in presence of several objects, the pose of the observer is optimized together with the relative position of the visible objects typically using temporal coherence constraints, i.e. objects projections in different images are expected to suggest the same act of motion. As the environment grows larger so does the number of required objects, thus yielding to computationally intensive algorithms.

To reduce the complexity Simon et al. [3] and Uematsu et al. [4] considers only planar reference objects. In this settings they can exploit both temporal and spatial coherence in the estimation, i.e. homographies between planes can be computed independently and added as additional constraints. This involves constructing at each frames a unified projective space and mapping all the planes to that space according to the computed homographies. The pose is subsequently computed using correspondences from that space and the image projections.

Nonetheless when several planar reference objects are also coplanar, the unified projective space can be profitably built in advance using image mosaicing techniques. As the cluster of objects becomes larger, using a mosaic as appearance model instead of a single shot, taken from larger distance or with shorter focal length, becomes more and more useful. In fact, the mosaic approach allows to maintain plenty of details that a single shot would miss.

We propose a practical approach that is suitable for environment where big planar structures are present. By mosaicing several planar objects during a training stage we shift off-line most part of the computation of the pose. At run-time, the algorithm simply determines the pose with respect to a single big reference object using approaches, such as [125], that are known to be fast and robust. This notably diminishes the on-line computational requirements and increases the accuracy of the estimated pose.

2 Methodology

The method is split up in two distinct stages. The first can be regarded as a training phase and it is performed off-line. It deals with the definition of a big planar reference object together with the construction of its appearance model, i.e. a mosaic of images that portray the planar structure. Several keypoints are extracted from the appearance model using the well known SIFT features

detector [6]. Metric measurements can be easily introduced in this framework by specifying the real world position of at least four non collinear points within the planar objects and computing the metric to projective homography accordingly.

The second stage performs on-line and addresses the estimation of the pose of the observer at a given instant using a set of points correspondences between the visible scene and the constructed appearance model. This stage is composed of a feature tracker that finds point matches and any chosen pose estimation algorithm based on point correspondences. The projection of virtual objects is easily accomplished once the pose is known.

2.1 Construction of the Appearance Model

The first stage concerns the construction of the big reference object and its appearance model from a collection of pictures using a mosaicing algorithm. The idea of using mosaics in augmented reality applications is not a novelty in itself. For instance, Dehais et al. [7] use mosaics to augment the scene with virtual objects. However, with their system the user is allowed to rotate only and both the training and the testing sequence must be captured from the same vantage point. The approach proposed by Liu et al. [8] is also based on image mosaicing, but it requires fiducial markers and the viewpoint is again allowed to rotate only. Instead, our method relies on natural markers present in the scene and allows for every kind of motion as long as a portion of the model is visible to the observer.

During a training stage we construct the appearance model using several views of a roughly planar structure in the scene. The transformations among the views are homographies as long as the observed subject is planar. The algorithm we use to mosaic images can be regarded as an iterative version of the pairwise DLT method described in [9]. From each pair of views we compute a set of point correspondences and fit the best homography in the least square sense. Then we repeat this procedure for all the pairs and concatenate the homographies. This can be seen as the common projective space computed by [4] when all the patterns are coplanar.

Instead of building a mosaic one might also capture the whole planar structure with a single shot taken from a larger distance or with a shorter focal length and then use such a shot as the appearance model. Indeed, this choice is potentially preferable when, given the resolution of the acquisition device, objects are as small as they can be captured by a single shot without losing too much information. In fact, in such a case objects are already registered with respect to each other and taking a picture is quicker than building a mosaic. However, in any application scenario the more appropriate approach should be identified carefully. In the experimental results section we propose a comparison between the two approaches considering two different case studies.

Finally, given the appearance model, the SIFT feature detector extracts a set of keypoints p_i from it. Extracted features that appear in the model but do not belong to the planar reference object are discarded using outlier removal techniques such as Ransac.

2.2 Pose Estimation and Augmentation

Pose estimation from point correspondences for calibrated perspective projection cameras has been extensively studied in literature. To demonstrate the effectiveness of our proposal we choose two well known algorithms that address the problem from very diverse points of view.

The pose estimation problem can be stated as that of estimating the rigid transformation, made up of a rotation matrix R and a translation vector t , that relates a set of noncollinear 3D coordinates of known reference points p_i with their corresponding normalized projections (u_i, v_i) so that:

$$\begin{aligned} u_i &= \frac{R^1 p_i + t_x}{R^3 p_i + t_z} \\ v_i &= \frac{R^2 p_i + t_y}{R^3 p_i + t_z} \end{aligned} \quad (1)$$

where $p_i = (x_i, y_i, z_i)$ are expressed in an object-centered frame, R is 3×3 orthonormal matrix and t is a 3×1 vector.

In these settings, the first algorithm we consider is that illustrated by Simon et al. [5], which has been considered for long the classical photogrammetric formulation. In practice they solve for the unknown pose by optimizing the following objective function:

$$\sum_i^N \left\| \left(\hat{u}_i - \frac{R^1 p_i + t_x}{R^3 p_i + t_z} \right), \left(\hat{v}_i - \frac{R^2 p_i + t_y}{R^3 p_i + t_z} \right) \right\|^2 \quad (2)$$

where \hat{u}_i, \hat{v}_i are measured image points. This computation minimizes the error distance among projections in the image space. In place of the sequential estimation proposed in their paper, we compute the pose of each frame with respect to our appearance model thus avoiding potential estimation drift issues.

From a theoretical viewpoint, an equivalent reformulation of the problem consists in estimating (R, t) that relates the known reference points p_i with the corresponding q_i so that:

$$q_i = Rp + t \quad (3)$$

where $p_i = (x_i, y_i, z_i)$ and $q_i = (x'_i, y'_i, z'_i)$ are expressed in an object-centered and camera-centered reference frame respectively. Based on this viewpoint, the second algorithm, proposed by Schweighofer et al. [1], aims at minimizing an object space distance error by means of the line-of-sight projection matrix \widehat{V}_i (for further details refer to [1]). This algorithm yields the best results according to a recent analysis of the state-of-the-art carried out in [2].

Once the pose is retrieved it is then possible to project 3D models in the image according to (R, t) and the known camera intrinsics.

3 Experimental Results

This section reports the performance of the considered algorithms in two different case studies. Performance are measured in terms of estimation steadiness and

smoothness. Under this perspective the most stable the estimated pose over time the better the algorithm. In the following we plot the recovered position of the camera center's coordinates $O^C = (O_X^C, O_Y^C, O_Z^C)$ expressed in the object-centered frame. Both algorithms are run twice on each sequence with different appearance models, the first time using a single image (Fig. 1 top), the second time using a mosaic (Fig. 1 bottom). All the frames used to build the models do not belong to the test sequences.

The two test sequences have been acquired by a freely moving observer using a webcam (Logitech Quick Cam Sphere). Each sequence is about 600 frames long and the images have a resolution of 640×480 pixels.



Fig. 1. Small (top) and large (bottom) appearance models

3.1 Aeronautical Servicing

The first case study is drawn from a collaborative research project addressing the application of Augmented Reality to the field of aeronautical servicing. The ultimate aim of the project is to equip engineers with see-through helmets by which a context-aware system will act as a virtual assistant providing information on the maintenance procedure in real-time using augmented reality. The sequence portraits the inside of a cockpit of a plane. Useful information in this context concerns the position of the most important switches and leverages as well as instructions on how to use them properly (refer to Fig. 3 for some examples).

In the upper row of Fig. 2 the position of O^C according to the pose estimated using a small appearance model is reported. While the pose is correct most of the time, the peaks in the plots denote that the estimation suffers from jitter. It is worth pointing out that both pose estimation methods are affected by these peaks approximately in the same way. Conversely, the plots in the lower row of Fig. 2 show that, when using the mosaic as appearance model, the estimated pose exhibits a much smoother behaviour and jitter is almost completely eliminated,

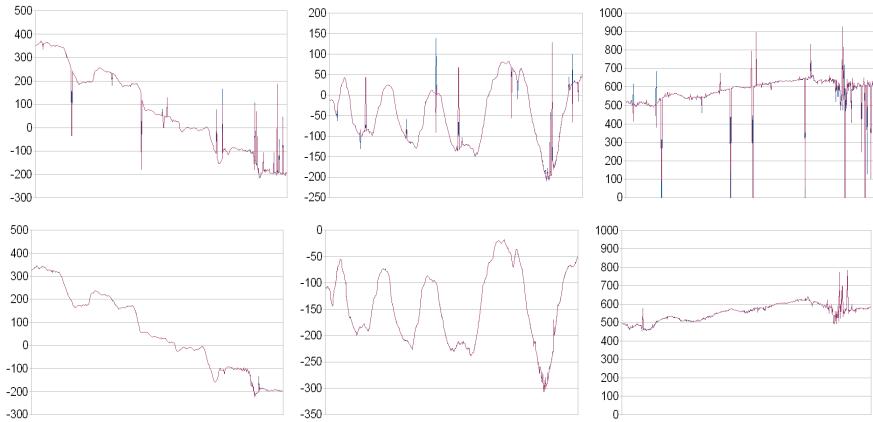


Fig. 2. Recovered camera center's coordinates using small (top) and large (bottom) appearance models: Schweighofer et al. (violet), Simon et al. (bordeaux). Left to right: O_X^C , O_Y^C , O_Z^C .



Fig. 3. Augmented cockpit sequence samples

with the exception of some creases regarding the z component. It is also worth noticing how the accuracy is not affected by the considerable lighting changes occurring in the scene, as shown by Fig. 3.

3.2 Cultural Heritage

The second case study concerns the field of advanced context-aware systems for delivering information to visitors of museums or archaeological sites. The considered sequence displays a showcase with Etruscan jewellery. Fig. 5 shows that the pose of the observer with respect to the showcase is accurately retrieved, as vouched by the coloured outlines superimposed on the borders of the shelves. Besides, additional context aware information is conveyed by highlighting the object that is likely to be the most important for the user given his position and orientation.

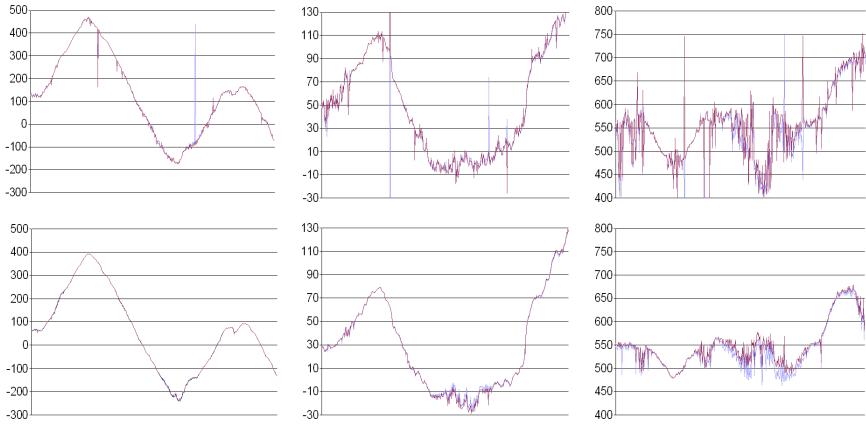


Fig. 4. Recovered camera center's coordinates using small (top) and large (bottom) appearance models: Schweigofer et al. (violet), Simon et al. (bordeaux). Left to right: O_X^C , O_Y^C , O_Z^C .

As before, the estimation using a small appearance model is quite good but suffers from jitter (as it can be seen in the upper row of Fig 4). When using the mosaic (lower row of Fig 4), jitter mostly disappears and, unlike previous experiment, the pose is smoother even when there are no macroscopic estimation error. Similarly to the cockpit sequence, the z component exhibits the worst reconstruction quality because the model is mostly observed from ahead with limited tilt angles, thus ill-conditioning the optimization procedure.



Fig. 5. Augmented jewel sequence samples

4 Conclusions

In this paper we have presented a practical approach to augmented reality that is suitable for environments where big planar objects are present. Instead of modeling the reference objects using a single image or a set of independent images, we propose to build a mosaic by registering several detailed views. The pose is then estimated from the correspondences between the actual frame and

the appearance model of the reference planar object using any chosen pose estimation algorithms. The experiments demonstrate that two very different pose estimation algorithms largely benefit from the proposed approach. In this sense our proposal can be thought as a preprocessing step able to improve the computational performance and accuracy of any pose estimation algorithms. The major limitation of the proposed approach is represented by the planar structure constraint.

References

1. Schweighofer, G., Pinz, A.: Robust pose estimation from a planar target. *IEEE Trans. Pattern Anal. Mach. Intell.* 28(12), 2024–2030 (2006)
2. Moreno-Noguer, F., Lepetit, V., Fua, P.: Accurate non-iterative $O(n)$ solution to the pnp problem. In: *IEEE Intl. Conf. on Computer Vision*, Rio de Janeiro, Brazil (October 2007)
3. Simon, G., Berger, M.-O.: Real time registration of known or recovered multi-planar structures: application to ar. In: *Proc. of BMVC* (2002)
4. Uematsu, Y., Saito, H.: Vision-based registration for augmented reality with integration of arbitrary multiple planes. In: *Proc. of ICIAP*, pp. 155–162 (2005)
5. Simon, G., Fitzgibbon, A.W., Zisserman, A.: Markerless tracking using planar structures in the scene. In: *Proc. of ISAR*, Munich, Germany, May–June 2000, pp. 120–128 (2000)
6. Lowe, D.: Distinctive image features from scale-invariant keypoints. *Intl. J. of Computer Vision* 60(2), 91–110 (2004)
7. Dehais, C., Douze, M., Morin, G., Charvillat, V.: Augmented reality through real-time tracking of video sequences using a panoramic view. In: *Proc. of ICPR*, pp. 995–998 (2004)
8. Liu, P., Sun, X., Georganas, N.D., Dubois, E.: Augmented reality: A novel approach for navigating in panorama-based virtual. In: *Proc. of HAVE* (2003) (unknown)
9. Hartley, R., Zisserman, A.: *Multiple View Geometry in Computer Vision*, 2nd edn. Cambridge University Press, Cambridge (2003)

Retracted: A Novel Text-Independent Speaker Verification System Using Ant Colony Optimization Algorithm

Shahla Nemati¹, Reza Boostani², and Mohammad Davarpanah Jazi¹

¹ Department of Electrical and Computer Engineering, Isfahan University of Technology(IUT), Isfahan 84156-83111, Iran

² Computer Sciences and Engineering, Campus No.2, Engineering School, Mollasadra St., Shiraz, Iran
s.nemati@ec.iut.ac.ir,
boostani@shirazu.ac.ir,
mdjazi@cc.iut.ac.ir

Abstract. Automatic speaker verification (ASV) has become increasingly desirable in recent years. This system in general, requires 20 to 40 features as input for satisfactory verification. In this paper, features size is reduced by Ant Colony Optimization (ACO) technique to increase the ASV performance. After feature reduction phase, feature vectors are applied to a Gaussian Mixture Model (GMM) which is a text-independent speaker verification Model. Experiments are conducted on a subset of TIMIT corpora. The results indicate that with the optimized feature set, the performance of the ASV system is improved. Moreover, the speed of verification is significantly increased because number of features is reduced over 73% which consequently decrease the complexity of our ASV system.

Keywords: Speaker Verification, Gaussian Mixture Model (GMM), Feature Selection, Ant Colony Optimization (ACO).

1 Introduction

Automatic Speaker Verification (ASV) refers to the task of verifying speaker's identity using speaker-specific information contained in speech signal. Speaker Verification methods are totally divided to text-dependent and text-independent applications. When the same text is used for both training and testing, the system is called to be text-dependent while for text-independent operation, the text used to train and test of the ASV system is completely unconstrained. Text independent speaker verification requires no restriction on the type of input speech. In contrast, Text independent speaker verification usually gives less performance than text dependent speaker verification, which requires test input to be the same sentence as training data [1].

Speech signals contain a huge amount of information and can be described as having a number of different levels of information. At the top level, we have lexical and syntactic features, below that are prosodic features, further below these

are phonetic features, and at the most basic level we have low-level acoustic features, which generally give information on the system that creates the sound, such as the speakers' vocal tract. Information solely about how the sound is produced (from low level acoustic features) should give enough information to identify accurately a speaker as this is naturally speaker dependent and independent of text [2].

Low level acoustic features also contain some redundant features, which can be eliminated using feature selection (FS) techniques. The objective of FS is to simplify a dataset by reducing its dimensionality and identifying relevant underlying features without sacrificing predictive accuracy. By doing that, it also reduces redundancy in the information provided by the selected features. Selected features should have high inter-class variance and low intra-class variability. Ideally they should also be as independent of each other as possible in order to minimize redundancy [3].

Feature selection has been rarely used in ASV systems. Day and Nandi [2] employed genetic programming (GP) for FS, also L plus-R minus feature selection algorithm is used by Pandit and Kittkr [4] for text-dependent speaker verification. Ant colony optimization (ACO) is a powerful method in many optimization methods [5] and has been employed here for feature selection in ASV systems. In this paper ACO algorithm has been applied to the problem of feature selection in ASV systems.

The rest of this paper is organized as follows. Section 2 presents a brief overview of ASV systems. Feature selection based ACO is described in Sections 3. Next section reports experimental results which include a brief discussion of the results obtained. Finally the conclusion and future research is offered in section 5.

2 An Overview of ASV Systems

The typical process in most proposed ASV systems involves: some form of pre-processing of the data (silence removal) and feature extraction, followed by some form of speaker modeling to estimate class dependent feature distributions (see Fig. 1). A comprehensive overview can be found in [6]. Adopting this strategy the ASV problem can be further divided into the two problem domains of:

- 1) Preprocessing, feature generation and selection.
- 2) Speaker modeling and matching.

2.1 Feature Extraction

Most previous works relied on the use of low-level acoustic features. Mel-frequency cepstral coefficients (MFCCs) have been particularly popular for ASV systems in recent years. This transform gives a highly compact representation of the spectral envelope of a sound. Many proposed systems have relied solely on these features and good results have been reported [1][7]. MFCCs have usually been associated

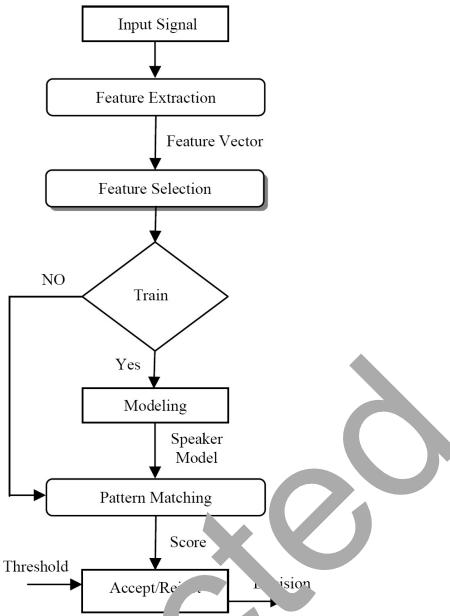


Fig. 1. Overview of the speaker verification process [8]

with speech analysis and processing, and can be used as a compact representation of the spectral envelope.

The process of MFCC analysis is summarized as follows [8]:

- ✓ Segmentation of a speech signal into frames with a fixed window size and a fixed shift period.
- ✓ For each frame:
 - Compute its fast Fourier transform (FFT).
 - Wrap the frequencies according to the Mel scale.
 - Compute the log of the magnitude.
 - Compute the discrete cosine transform (DCT).

2.2 Speaker Modeling

The speaker modeling stage of the process varies more in the literature. The purpose of speaker modeling is characterizing an individual which is enrolled into an ASV system with the aim of defining a model (usually feature distribution values). The two most popular methods in previous works are Gaussian mixture models (GMM) [7] and vector quantization (VQ) [9]. Other techniques such as mono-Gaussian models [10], decision trees [11], SVM [12], and ANN [13] have also been applied. In this paper GMM is used for speaker modeling.

The GMM method is a parametric method that consists of M Gaussian distributions parameterized by their priori probabilities, mean vectors and covariance

matrices. The parameters are typically estimated by maximum likelihood (ML) estimation [7].

Let $X = \{x_1, x_2, \dots, x_T\}$ be a set of T vectors, each of which is a d-dimensional feature vector extracted by digital speech signal processing. Since the distribution of these vectors is unknown, it is approximately modeled by a mixture of Gaussian densities, which is a weighted sum of M component densities, given by the equation:

$$p(x_t|\lambda) = \sum_{i=1}^M w_i N(x_t, \mu_i, \Sigma_i) \quad (1)$$

where λ denotes a prototype consisting of a set of model parameters $\lambda = \{w_i, \mu_i, \Sigma_i\}$, w_i are the mixture weights and $N(x_t, \mu_i, \Sigma_i)$ are the d-variate Gaussian component densities with mean vectors μ_i and covariance matrices Σ_i

$$N(x_t, \mu_i, \Sigma_i) = \frac{\exp\{-\frac{1}{2}(x_t - \mu_i)' \Sigma_i^{-1} (x_t - \mu_i)\}}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \quad (2)$$

In training the GMM, these parameters are estimated such that they are fitted with the distribution of the training vectors. The most widely used training method is the ML estimation. For a sequence of training vectors X , the likelihood of the GMM is:

$$p(X|\lambda) = \prod_{t=1}^T p(x_t|\lambda) \quad (3)$$

The classification for speaker verification may be performed based on the log likelihood ratio (LLR). Given the observation X , the LLR is defined as:

$$LLR(X) = \log p(X|\lambda) \quad (4)$$

Let Θ be a given threshold, associated with the claimed speaker model λ . A discriminant function $LLR(X)$ for the unknown X and model λ used to reject or accept the claimed speaker is as follows:

$$LLR(X) \begin{cases} \geq \Theta & \text{Accept} \\ < \Theta & \text{Reject} \end{cases} \quad (5)$$

Where threshold Θ is taken as an equal error rate (EER) threshold [8].

3 Application of ACO for Feature Selection

In the early 1990s, ant colony optimization (ACO) was introduced by M. Dorigo and colleagues as a novel nature-inspired meta-heuristic for the solution of hard combinatorial optimization (CO) problems [5]. ACO belongs to the class of meta-heuristics, which includes approximate algorithms used to obtain good enough solutions to hard CO problems in a reasonable amount of computation time.

The paradigm is based on the observation made by ethologists about the medium used by ants to communicate information regarding shortest paths to

food by means of pheromone trails. A moving ant lays some pheromone on the ground, thus, a path is made by a trail of this substance. While an isolated ant moves practically at random (exploration), an ant encountering a previously laid trail can detect it and decide with high probability to follow it and consequently reinforce the trail with its own pheromone (exploitation) [5].

The feature selection task is reformulated based on the ACO which requires a problem to be represented as a graph. Here nodes represent features, with the edges between them denoting the choice of the next feature. The search for the optimal feature subset is then an ant traversal through the graph where a minimum number of nodes are visited that satisfies the traversal stopping criterion. The heuristic desirability of traversal and edge pheromone levels are combined to form the so-called probabilistic transition rule, denoting the probability of ant k at feature i choosing to travel to feature j at time t [14]:

$$P_{ij}^k(t) = \begin{cases} \frac{[\tau_{ij}(t)]^\alpha \cdot [\eta_{ij}]^\beta}{\sum_{l \in J_i^k} [\tau_{il}(t)]^\alpha \cdot [\eta_{il}]^\beta} & j \in J_i^k \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Where, η_{ij} is the heuristic desirability of choosing feature j when at feature i (η_{ij} is optional but often needed for achieving a high algorithm performance), J_i^k is the set of neighbor nodes of node i which have not yet been visited by the ant k . $\alpha > 0, \beta > 0$ are two parameters that determine the relative importance of the pheromone value and heuristic information (the choice of α, β is determined experimentally) and $\tau_{ij}(t)$ is the amount of virtual pheromone on edge (i,j) . The pheromone on each edge is updated according to the following formula:

$$\tau_{ij}(t+1) = (1 - \rho) \cdot \tau_{ij}(t) + \sum_{k=1}^m \Delta_{ij}^k(t) \quad (7)$$

where:

$$\Delta_{ij}^k(t) = \begin{cases} \gamma'(S^k)/|S^k| & \text{if } (i, j) \in S^k \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

The value $0 \leq \rho \leq 1$ is decay constant used to simulate the evaporation of the pheromone, S^k is the feature subset found by ant k . The pheromone is updated according to both the measure of the "goodness" of the ant's feature subset (γ') and the size of the subset itself. By this definition, all ants can update the pheromone. The overall process of ACO feature selection for ASV is shown in Fig. 2.

4 Experimental Results

4.1 TIMIT Dataset

The TIMIT corpora [15] is used in this paper. This corpus contains 630 speakers (438 male and 192 female) representing 8 major dialect regions of the United States, each speaking ten sentences. There are two sentences that are spoken by

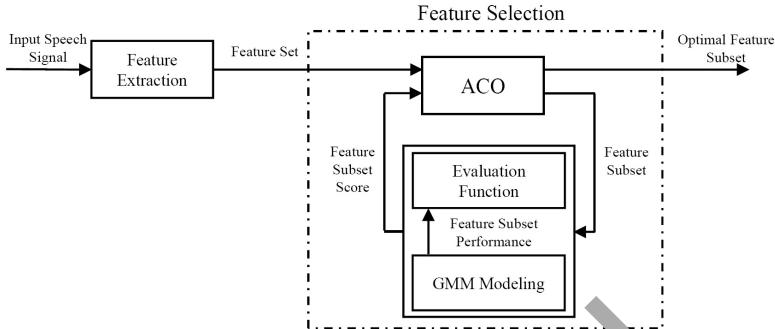


Fig. 2. Overall process of ACO feature selection for ASV

all speakers and the remaining eight are selected randomly from a large database. The speech signal is recorded through a high quality microphone with a sampling frequency of 16 kHz in a quiet environment, with no session interval between recordings.

Eight sentences (SX, SI) were used to develop each speaker model, and the remaining 2 SA sentences were used to test each speaker. The 35 speakers included in both the test and train directories were used during the TIMIT(35) trials.

4.2 Evaluation Measure

The evaluation of the speaker verification system is based on detection error tradeoff (DET) curves, which show the tradeoff between false alarm (FA) and false rejection (FR) errors. Typically equal error rate (EER), which is the point on the curve where $FA = FR$, is chosen as evaluation measure. We also used detection cost function (DCF) defined as [7]:

$$DCF = C_{miss} \cdot E_{miss} \cdot P_{target} + C_{fa} \cdot E_{fa} \cdot (1 - P_{target}) \quad (9)$$

where P_{target} is the priori probability of target tests with $P_{target} = 0.01$, E_{miss} and E_{fa} are false rejection rate and false acceptance rate respectively at a operating point and the specific cost factors $C_{miss} = 10$ and $C_{fa} = 1$. Hence, the point of interest is shifted towards low FA rates.

4.3 Experimental Setup

Experiments were conducted on a subset of TIMIT corpora consists of 22 male and 13 female speakers of different accent that were selected randomly. Data were processed in 30 ms frames (480 samples) with 50% overlaps. Frames were segmented by Hamming window and pre-emphasized with $\alpha = 0.97$ to compensate the effect of microphone's low pass filter. At first, for each frame MFCCs were extracted from silence removed data. Moreover, delta coefficients were calculated based on the MFCCs and appended to them. Furthermore, two energy terms were also extracted to form input vectors. ACO-based feature selection

Table 1. Selected Features of ACO-GMM

Method	Number of Original Features	Number of Selected Features	Selected Features
ACO-GMM (32)	26	7	1,2,3,4,6,9,12
ACO-GMM (32)	42	7	1,4,5,7,9,18,24
ACO-GMM (64)	26	7	1,3,4,6,7,8,10
ACO-GMM (64)	42	6	1,4,5,6,8,13

Table 2. Equal Error Rate (EER) Results (%) and DCF for GMM and ACO-GMM

Number of MFCCs	GMM-32		ACO-GMM-32		GMM-64		ACO-GMM-64	
	EER	DCF	EER	DCF	EER	DCF	EER	DCF
26	6.79	0.0689	3.76	0.0388	6.99	0.0771	7	0.0691
42	9.65	0.0937	8.16	0.0615	8.64	0.0841	9	0.0943

was applied to input vectors which was described earlier. Then, verification process was performed using the GMM approach. The performance criterion is due to EER and DCF according to an adopted decision threshold strategy.

Table 1 shows the number of selected features and selected features which were chosen by ACO. As we can see in table 1, ACO can degrade dimensionality of features over 73%. EER and DCF for GMM and ACO-GMM with different number of Gaussian (32, 64) were shown in Table 2.

DET curves for GMM and ACO-GMM with 32 Gaussian are shown in Figure 3, and those for 64 Gaussian are shown in Figure 4. From the results, it can be seen that ACO-GMM yields a significant improvement in speed than the baseline GMM approach. The improvement is due to the selection of optimal feature set by ACO algorithm.

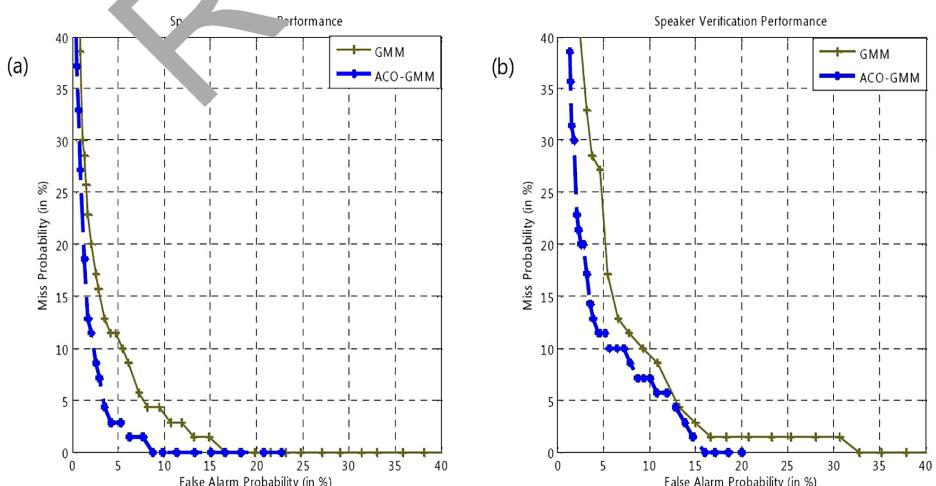


Fig. 3. DET curves for GMM and ACO-GMM with 32 Gaussians (a)original features 26 (b) original features 42

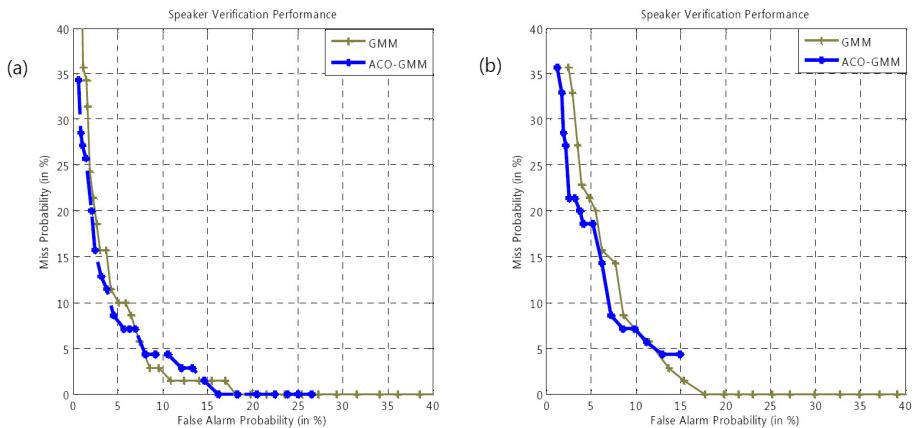


Fig. 4. DET curves for GMM and ACO-GMM with 64 Gaussians (a)original features 26 (b) original features 42

5 Conclusion and Future Research

In this paper, we have addressed the problem of optimizing the acoustic feature set by ACO technique for text-independent speaker verification system based on GMM. ACO selected the relevant features among all Mel-Cepstrum coefficients in order to increase the performance of our ASV system. The experimental results on subset of TIMIT database showed that ACO is able to select the most informative features without loosing the performance. The feature vectors size reduced over 73% which led to a less complexity of our ASV system. Moreover, verification process in the test phase speeds up because less complexity is achieved by the proposed system in comparison with current ASV systems.

For future work, the authors plan to investigate the performance of proposed ASV system by taking advantage of using GMM-UBM and other models instead of GMM. Finally, another research direction will involve experiments with different datasets.

Acknowledgments. The authors wish to thank the Office of Graduate studies of the Isfahan University of Technology for their support.

References

1. Xiang, B., Berger, T.: Efficient Text-Independent Speaker Verification with Structural Gaussian Mixture Models and Neural Network. *IEEE Transactions On Speech And Audio Processing* 11(5) (2003)
2. Day, P., Nandi, A.K.: Robust Text-Independent Speaker Verification Using Genetic Programming. *IEEE Transactions On Audio, Speech, And Language Processing* 15(1), 285–295 (2007)

3. Jensen, R.: Combining rough and fuzzy sets for feature selection. Ph.D. Thesis, University of Edinburgh (2005)
4. Pandit, M., Kittkr, J.: Feature Selection for a DTW-Based Speaker Verification System, pp. 796–799. IEEE, Los Alamitos (1998)
5. Dorigo, M., Caro, G.D.: Ant Colony Optimization: A New Meta-heuristic. In: Proceeding of the Congress on Evolutionary Computing (1999)
6. Campbell, J.P.: Speaker recognition: a tutorial. Proc. IEEE 85(9), 1437–1462 (1997)
7. Reynolds, D.A., Rose, R.C.: Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions On Speech And Audio Processing 3(1), 72–83 (1995)
8. Cheung-chi, L.: GMM-Based Speaker Recognition for Mobile Embedded Systems, Ph.D. Thesis, University of Hong Kong (2004)
9. Linde, Y., Buzo, A., Gray, R.M.: An Algorithm for Vector Quantizer Design. IEEE Trans. Comm. 28, 84–95 (1980)
10. Bimbot, F., Magrin-Chagnolleau, I., Mathan, L.: Second-order statistical measures for text-independent speaker identification. Speech Commun 17(12), 177–192 (1995)
11. Navratil, J., Jin, Q., Andrews, W., Campbell, J.: Phonetic speaker recognition using maximum-likelihood binary-decision tree models. In: Proc. ICASSP, Hong Kong (2003)
12. Wan, V.: Speaker Verification Using Support Vector Machines, Ph.D. dissertation, Univ. Sheffield, U.K (2003)
13. Wouhaybi, R., Al-Alaou, M.A.: Comparison of neural networks for speaker recognition. In: Proc. 6th IEEE Int. Conf. Electronics, Circuits Systems (ICECS), vol. 1, pp. 125–128 (1999)
14. Kanan, H.R., Faez, K., Hosseinzadeh, M.: Face Recognition System Using Ant Colony Optimization-Based Selected Features. In: CISDA 2007. Proceeding of the First IEEE Symposium on Computational Intelligence in Security and Defense Applications, pp. 57–62. IEEE Press, Los Alamitos (2007)
15. Garofolo, J., et al.: DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. National Institute of Standards and Technology (1990)

Image Retrieval: Color and Texture Combining Based on Query-Image*

Ilya Markov¹ and Natalia Vassilieva²

¹ Saint-Petersburg State University, Russia

ilya.markov@gmail.com

² HP Labs, Saint-Petersburg, Russia

nvassilieva@hp.com

Abstract. It is a common way to process different image features independently in order to measure similarity between images. Color and texture are the common ones to use for searching in natural images. In [10] a technique to combine color and texture features based on a particular query-image in order to improve retrieval efficiency was proposed. Weighted linear combination of color and texture metrics was considered as a mixed-metrics. In this paper the mixed-metrics with different weights are compared to pure color and texture metrics and widely used CombMNZ data fusion algorithm. Experiments show that proposed metrics outperform CombMNZ method in some cases, and have close results in others.

Keywords: Content-Based Image Retrieval, Mixed-Metrics, Data Fusion.

1 Introduction

Color and texture are the common features which are used for searching in natural images with heterogeneous content. Heterogeneous content means that considered image collection has no common subject or properties. In case collection has one common theme (for example, facial collection, collection of finger prints, medical shots collection), it might be possible to use some particular set of features adjusted to the theme, which might be more effective.

Common approach in image retrieval is to measure similarity between images based on different features independently and than combine these results together in order to get final result. While there are a lot of studies on different low-level image features analysis [12], not so many researches consider a task of combining various similarity measures based on different image features.

In image retrieval linear combination is commonly used to combine multiple searches in order to get the final result [12][5]. The main reason for that is the simplicity of the algorithm. Text retrieval research domain has longer history and more experience. It is possible to borrow some methods from that area and apply them successfully to image retrieval. CombMNZ [4] is considered to be one of the best data fusion algorithms for combining multiple searches in text retrieval

* This work was partially supported by RFBR (grant 07-07-00268a).

In [17] we showed that it can be used in image retrieval also: CombMNZ outperforms linear combination in most cases.

In [10] we proposed a technique to combine color and texture metrics taking into account a particular query-image without interaction between system and human. Weighed linear combination of color and texture metrics (mixed-metrics) is considered as a fusion function. Our approach is based on the hypothesis, proposed and proved in [10], that there are optimal weights to combine color and texture metrics for every query-image and these weights are unique for a given query. By using these optimal weights one can improve retrieval results. We showed that it is always possible to mark out the best mixed-metrics for every group of similar images (and thus for every query-image). In other words, we proposed adaptive fusion algorithm without using relevance feedback.

It is possible that one can obtain better results by utilizing relevance feedback algorithms, which are sometimes used to obtain optimal coefficients for every particular query to combine different similarity measures [1]. But, on one hand, not all image retrieval systems have relevance feedback implementation and, on the other, our method can improve even those which have. When relevance feedback is used, improvement can be achieved starting from the second iteration. Our method can be used to get more precise retrieval results on the first iteration, when no feedback from user is yet available.

In this study we continue to investigate the approach described above and compare it to CombMNZ [4] algorithm. Experimental results show that mixed-metrics outperform CombMNZ method in some cases, and have close results in others. In average mixed-metrics slightly outperform CombNMZ algorithm: average precision of mixed-metrics search is 42.76% and the same for CombMNZ is 39.68%.

2 Related Works

Many researchers showed that it is necessary to combine various features for effective image retrieval. At the same time not enough attention is paid to the particular fusion methods. A number of works dedicated to similarity measures combining for image retrieval task is relatively small.

In [3] authors examine an application of a fuzzy logic approach to the fusion of image features. While it is a promising technique, there is no similarity measure proposed for observed fused feature and no experimental or other results are shown that can prove an efficiency of this approach.

Common solution is to fuse similarity measures calculated based on different features but not the features themselves. Linear combination of multiple similarity measures is usually treated as an aggregate measure (in [5] for instance).

Common data fusion algorithms like CombSUM, CombMNZ [4] and others [8,7] are widely used in text retrieval. The same algorithms can be applied to image retrieval domain.

CombMNZ is considered to be one of the best data fusion algorithms. It performs as follows. Element in the result ranked-list gets rank equaled to the sum of all its ranks in fused lists divided by the number of lists in which this element exists with non-zero rank:

$$\begin{aligned} rank_{result}(obj) &= \sum_{fused\ lists} rank_{list}(obj) \cdot nz, \quad \forall obj \in image\ collection, \\ &\text{where } nz = \sum_{fused\ lists} (rank_{list}(obj) = 0 ? 0 : 1). \end{aligned}$$

This algorithm is simple to use and outperforms other data fusion methods [7].

In [17] we proposed our own data fusion method "Weighted Total with Gravitation Function" (WTGF) and compared it to CombMNZ, applied to the image retrieval domain. WTGF function satisfies various criteria like symmetry, monotonicity and so called "cone rules". Experimental results showed that WTGF outperforms CombMNZ in case there are multiple inputs of non-equal reliability (we can trust to one input more than to others) and inputs do not overlap much. In case information about element ranks is not trusted (all inputs have the same reliability) and inputs overlap a lot, CombMNZ outperforms WTGF.

Combination of search results obtained by using color and texture features is the second case. Therefore we compare mixed-metrics to CombMNZ algorithm in this work.

3 Mixed-Metrics

In [10] we proposed a technique to combine image similarity measures which takes into account a particular query-image. We introduced mixed-metrics obtained from color and texture metrics (C and T respectively) by using their weighted linear combination $a \cdot C + (1 - a) \cdot T$, where a is a varying coefficient which depends on a query-image.

We stated and proved the hypothesis that optimal value of a is the same for similar query-images. It means that in order to perform a search by using mixed-metrics one should go through the following steps. Entire image collection on which a search to be performed should be prepared as follows. Get some relatively small training set of images representing the whole collection and divide it into groups of similar images. Then calculate an average precision for every group for different values of a (varying from 0 to 1 with predefined step) applied to mixed-metrics. An average precision for a group is calculated based on retrievals when group's images are used as queries. Finally, select optimal coefficient a based on precision values and calculate "average" features for each group. One of the main ideas here is that this preparation should be done only once for the collection. It is also possible to use the same training set for several collections in case it represents all of them well.

To perform a search itself, one should classify query-image to one of the groups of the training set. After classification a search can be performed by using mixed-metrics of the group which query-image belongs to.

4 Color and Texture Features Selection

We use moment based color distribution features and color metrics from [15]. This approach is more robust in matching colors than the one based on classic color histograms [14]. In [15] color is represented with its mean for every channel and the covariance matrix of the channels' distributions. Minimal amount of spatial information is encoded into color index: each image is divided into five partially overlapping fuzzy regions. Feature vector is calculated for every region. Weighted Manhattan distance is used as a similarity function.

While ICA filters are more natural comparing to Gabor filters, we use convolutions of image with ICA filters as a texture feature and Kullback-Leibler divergence as a texture metrics [2]. Texture features built by using Gabor filters are one of the most popular approaches for texture. It was shown that Gabor-based features outperform other texture features in a query-by-example approach to image retrieval [9][6]. In [13] it was shown that ICA-filters perform better in classification task, therefore we can assume that they are better in retrieval task too.

5 Experiment

Experimental image database consists of 650 images from Corel Photo Set collection. It is divided into 9 groups based on images content by 2 experts. Result groups are: City, Clouds, Coastal landscapes, Contemporary buildings, Fields, Lakes, People, Rocks and Trees. This set of images can be considered as a training set for some larger collection.

For every image in the database color and texture features are extracted and for every pair of images color and texture distances are computed. Distance values are normalized according to the following rule:

$$distance_{result}(image) = (distance(image) - Average)/Deviation.$$

Therefore distributions of color and texture distances have the same *Average* and *Deviation*.

CombMNZ and several mixed-metrics are evaluated in our experiment. Participated mixed-metrics are: $a \cdot C + (1 - a) \cdot T$, where a varies from 0 to 1 with 0.1 step. Mixed-metrics with $a = 0$ is a pure texture metrics and with $a = 1$ is a pure color one.

To estimate retrieval efficiency we use average precision at N measure - a common one in information retrieval. Precision at N is a percentage of relevant objects among first N retrieved. To obtain average precision at N for all fusion methods the following procedure is performed. Every image in the database is used as a query. One search per every query and fusion method is run and precision at N for each run is calculated ($N = 1..30$). Images from the same group as the query-image are treated as relevant, while others are not. Average precision at N is calculated for every group of similar images and every fusion method.

6 Results Analysis

Summary results are shown in table 1 for the following search methods: pure color- and texture-based (ColorMoments and ICA respectively), CombMNZ and best mixed-metrics-based (in accordance with the particular group). For every group average precision at N is calculated for every method for different N.

Table 1. Best mixed-metrics precision compared to other metrics precisions for every group

Group	Search Algorithm	Average precision at N (%)					
		N=5	N=10	N=15	N=20	N=25	N=30
City	Color Moments	24	24	21	19	19	18
	ICA	31	27	26	26	23	22
	CombMNZ	39	32	27	24	22	20
	Mixed-Metrics (0.2)	44	38	35	33	30	29
Clouds	Color Moments	83	81	80	77	76	74
	ICA	68	62	57	52	49	47
	CombMNZ	82	80	79	78	77	75
	Mixed-Metrics (0.5)	86	81	79	78	76	75
Coastal Landscapes	Color Moments	36	34	33	32	31	31
	ICA	20	17	16	16	15	15
	CombMNZ	36	33	33	33	32	31
	Mixed-Metrics (0.8)	36	34	33	31	32	30
Contemp. Buildings	Color Moments	30	28	28	28	27	27
	ICA	22	21	21	19	19	18
	CombMNZ	30	29	31	30	30	30
	Mixed-Metrics (0.2)	29	32	32	33	32	32
Fields	Color Moments	51	50	48	45	44	43
	ICA	38	35	33	32	32	31
	CombMNZ	49	45	44	44	44	44
	Mixed-Metrics (1.0)	51	50	48	45	44	43
Lakes	Color Moments	45	43	43	42	41	40
	ICA	24	23	23	22	23	23
	CombMNZ	41	41	39	39	38	37
	Mixed-Metrics (0.8)	46	46	45	43	42	41
People	Color Moments	32	33	30	28	27	-
	ICA	26	20	20	18	17	-
	CombMNZ	40	34	31	29	27	-
	Mixed-Metrics (0.7)	42	39	39	36	35	-
Rocks	Color Moments	42	36	34	32	-	-
	ICA	6	5	4	5	-	-
	CombMNZ	46	36	33	28	-	-
	Mixed-Metrics (1.0)	42	36	34	32	-	-
Trees	Color Moments	35	33	32	32	32	32
	ICA	27	24	23	23	21	21
	CombMNZ	43	40	35	34	32	32
	Mixed-Metrics (0.2)	45	44	42	40	40	39

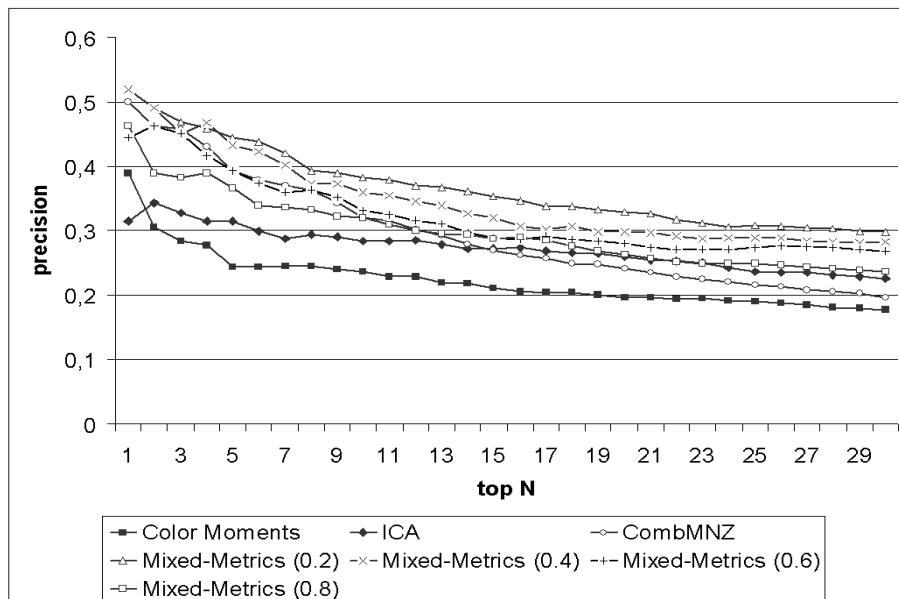


Fig. 1. Precision/Top N dependencies for different searches for group "City"

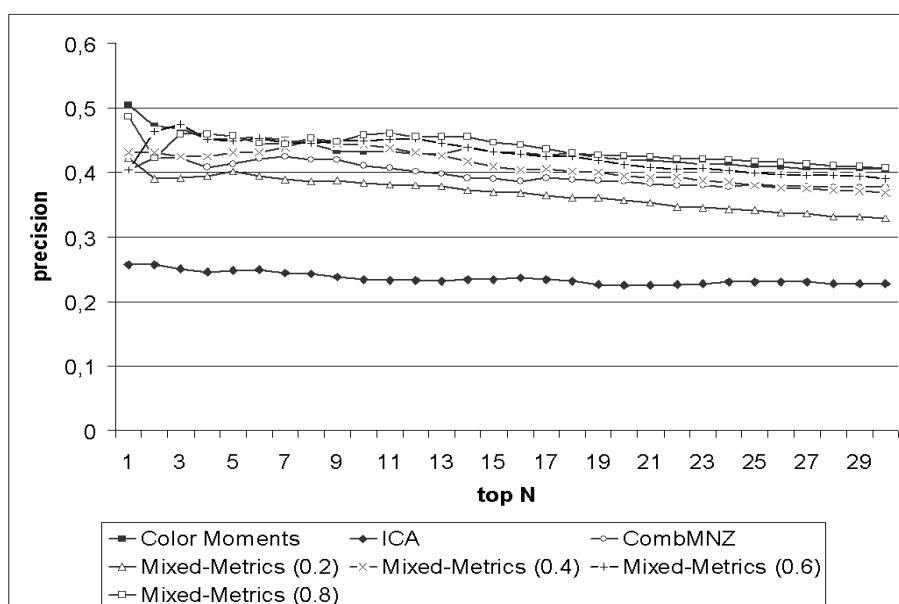


Fig. 2. Precision/Top N dependencies for different searches for group "Lakes"

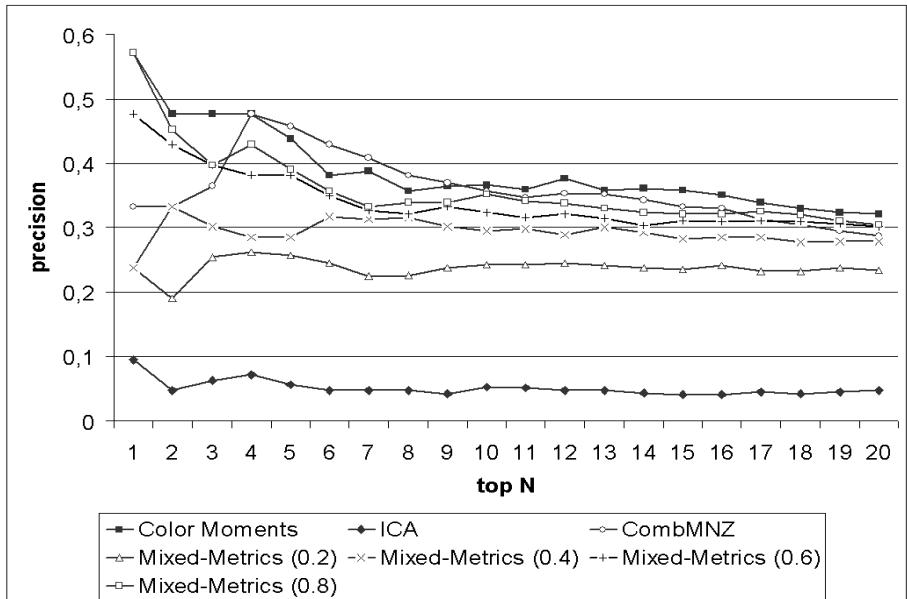


Fig. 3. Precision/Top N dependencies for different searches for group "Rocks"

Results show that in most cases a search with best group mixed-metrics has greater precision compared to runs with CombMNZ. For other groups ("Clouds", "Coastal landscapes" and "Rocks") results of both methods are very close to each other. Average precision of mixed-metrics search among all images of the database is 42.76% and the same for CombMNZ is 39.68%.

Let us discuss results for some groups in detail. Result charts for these groups ("City", "Lakes" and "Rocks") are shown on fig. 1, 2 and 3 respectively.

The precision/top N dependencies for different searches for group "City" are shown on fig. 1. Texture search gives more precise result than color one and therefore texture feature is more important for this group. Search with $0.2 \cdot C + 0.8 \cdot T$ mixed-metrics gives the best result and also proves the above statement. Results for other mixed-metrics searches show that precision decreases when mixed-metrics moves from texture to color. Precision of CombMNZ algorithm decreases faster than precision of color and texture searches. For first 10 positions it loses only 5% compared to $0.2 \cdot C + 0.8 \cdot T$ mixed-metrics. And for 30th position it loses 10%.

The inverse situation can be seen on fig. 2 for group "Lakes". Color feature is more important here and precision decreases when mixed-metrics moves from color to texture. Best result is obtained by using $0.8 \cdot C + 0.2 \cdot T$ mixed-metrics.

Group "Rocks" on fig. 3 is the case when search with pure color metrics and CombMNZ algorithm give more precise results. Pure color metrics can be treated as $1.0 \cdot C + 0.0 \cdot T$ mixed-metrics here.

7 Conclusions and Further Work

Experiments show that mixed-metrics improve retrieval results compared to pure color and pure texture metrics. This result proves common observation that combined search with several features gives better results than individual searchers. Moreover mixed-metrics outperform CombMNZ data fusion algorithm in some cases and give close results in others.

As it was mentioned in section 3 in order to perform retrieval using the optimal mixed-metrics a query-image should be classified to one of the groups established in training set. Therefore the next step of our research is to provide an efficient and effective classification algorithm for this task. While classification should be performed in real time during retrieval process it should be as fast as possible, therefore it should be simple enough and involve just a few computations. For this reason, many well-known classification algorithms cannot be used in our environment.

Possible solution here is to compute common color and texture features for obtained groups. Then the classification can be done through comparing query-image features to groups' common features. It is important that groups' common features should differ one from another. Otherwise it is impossible to perform classification task and obtain optimal mixed-metrics for any query-image.

Acknowledgments. This work was partially supported by RFBR (grant 07-07-00268a).

References

1. Aksoy, S., Haralick, R.M., Cheikh, F.A., Gabbouj, M.: A Weighted Distance Approach to Relevance Feedback. In: 15th International Conference on Pattern Recognition, vol. 4, pp. 812–815 (2000)
2. Borgne, H., Guerin-Dugue, A., Antoniadis, A.: Representation of images for classification with independent features. Pattern Recognition Letters 25, 141–154 (2004)
3. Deer, P.J., Eklund, P.W.: On the Fusion of Image Features, <http://citeseer.ist.psu.edu/162546.html>
4. Fox, E.A., Shaw, J.A.: Combination of multiple searches. TREC 2, 243–249 (1994)
5. Guerin-Dugue, A., Ayache, S., Berrut, C.: Image retrieval: a first step for a human centered approach. In: Joint Conference of ICI, CSP and PRCM, vol. 1, pp. 21 – 25 (2003)
6. Howarth, P., Rueger, S.: Robust texture features for still image retrieval. In: IEE Proc. of Vision, Image and Signal Processing, vol. 152(6), pp. 868–874 (2005)
7. Lee, J.H.: Analyses of Multiple Evidence Combination. In: ACM-SIGIR, USA, pp. 267–276 (1997)
8. Lilis, D., Toolan, F., Mur, A., Peng, L., Colier, R., Dunnion, J.: Probability-Based Fusion of Information Retrieval Result Sets. J. Artif. Intell. Rev. 25(1-2), 179–191 (2006)
9. Manjunath, B.S., Ma, W.Y.: Texture features for browsing and retrieval of image data. IEEE Transactions on Pattern Analysis and Machine Intelligence 18(8), 837–842 (1996)

10. Markov, I., Vassilieva, N., Yaremchuk, A.: Image retrieval. Optimal weights for color and texture features combining based on query object. In: Proc. of RCDL, Russia, pp. 195–200 (2007)
11. Montague, M., Aslam, J.A.: Relevance Score Normalization for Metasearch. In: ACM Conference on Information and Knowledge Management, pp. 427–433 (2001)
12. Rui, Y., Huang, T.S., Chang, S.-F.: Image Retrieval: Past, Present and Future. In: International Symposium on Multimedia Information Processing (1997)
13. Snitkowska, E., Kasprzak, W.: Independent Component Analysis of Textures in Angiography Images. Computational Imaging and Vision 32, 367–372 (2006)
14. Stricker, M., Dimai, A.: Color Indexing with Weak Spatial Constraints. In: Storage and Retrieval for Image and Video Databases (SPIE), pp. 26–40 (1996)
15. Stricker, M., Dimai, A.: Spectral Covariance and Fuzzy Regions for Image Indexing. In: Machine Vision and Applications, vol. 10, pp. 66–73 (1997)
16. Swain, M., Ballard, D.: Color Indexing. International Journal of Computer Vision 7(1), 11–32 (1991)
17. Vassilieva, N., Dolnik, A., Markov, I.: Image retrieval. Fusion of the result sets retrieved by using different image characteristics. Internet-mathematics Collection, 46–55 (2007)

Markerless Outdoor Localisation Based on SIFT Descriptors for Mobile Applications

Frank Lorenz Wendt, Stéphane Bres, Bruno Tellez, and Robert Laurini

LIRIS UMR 5205 CNRS, INSA-Lyon, F-69621, France
{lorenz.wendt, stephane.bres, bruno.tellez,
robert.laurini}@liris.cnrs.fr
<http://liris.cnrs.fr/>

Abstract. This study proposes augmented reality from mobile devices based on SIFT (Scale Invariant Feature Transform) features for markerless outdoor augmented reality application. The proposed application is navigation help in a city. These SIFT features are projected on a digital model of the building façades of the square to obtain 3D co-ordinates for each feature point. The algorithms implemented calculate the camera pose for frame of a video from 3D-2D point correspondences between features extracted in the current video frame and points in the reference dataset. The algorithms were successfully tested on video films of city squares. Although they do not operate in real-time, they are capable of a correct pose estimation and projection of artificial data into the scene. In case of a loss of track, the algorithms recover automatically. The study shows the potential of SIFT features for purely image based markerless outdoor augmented reality applications. This study takes place in the MoSAIC¹ project.

Keywords: Content-based image retrieval, image matching, augmented reality, SIFT, building recognition, pose estimation.

1 Introduction

The Internet has become the most important tool for information access and distribution. However, up to now, Internet access was restricted to stationary computers based in an office or at home, and linked to the Internet via cable. Existing options to access the Internet from a mobile client have been either restricted to selected places with a WLAN hotspot, or have provided only limited transfer rates via a mobile telephone network, or have been too costly to gain a significant market acceptance. In future, this is changing. A new network infrastructure is being built up, that grants mobile Internet access at transfer speeds comparable to home-based solutions. Mobile telephone manufacturers are developing optimized mobile handsets with more processing power and bigger displays, and Internet search engines offer services optimized for a mobile use. Moreover, nowadays standard equipment of any mobile device includes camera, GPS, and probably more interesting things in the future, which can be used as

¹ MoSAIC: Mobile Search and Annotation using Images in Context, ICT ASIA project (Ministère des Affaires Etrangères - MAE, France), 2006-2008.

new input medium to formulate search engine queries instead of using the more tedious textual input.

This study proposes a mobile, markerless augmented reality as a solution for a convenient and intuitive way to launch Internet queries with little or no need to enter a query as text, and to display query results in a simple and clear manner. The user would simply point with his camera at the building or object he is interested in (a restaurant, a theatre, a statue...), and the system would query a database or the Internet about the object of choice, and display the results in the live video of the object the user is just filming. Concepts for multimodal input options as proposed for example by Lim et al. [7] where the user takes a photo of the object of interest with his camera-equipped telephone are naturally included within this framework. Obviously, such a tool would require a combination of solutions to work properly – this study focuses only on the augmented reality solution itself, identifying the object in the view of the camera, and tracking it through the sequence of video images.

This system would be equally suited for navigation, tourism, gaming or advertisement.

Section 2 will give an overview to solutions to this problem proposed in literature. Section 3 cover our actual research and section 4 describe the results we have through a prototype of an augmented reality. Section 5 concludes the paper and raises further research questions.

2 Some Previous Works

In this study, an outdoor use of a hand-held augmented reality is proposed. This setting has some distinctive difficulties to overcome compared to an indoor application in a controlled environment. These difficulties can be summarized as follows : in an urban outdoor scene, abundant moving objects like cars and people can disturb the tracking process. The result is a camera movement that follows the cars, and not, as intended, the buildings. Likewise may a scene be partially occluded by objects that have not been at that position when the scene was modelled. The algorithm should nevertheless be able to recover the camera position from the remaining information. Plants are a further difficulty: They change their appearance over time, and can therefore not be used as “landmarks”. The fourth error source is lighting: in an outdoor application, lighting can not be controlled. Therefore, the visual appearance of objects can change considerably, making tracking difficult. Shadows may also produce edges that distract tracking.

2.1 Proposed Solutions

Several studies support or replace the visual tracking by additional sensors like gyroscopes, accelerometers, GPS modules and digital compasses to overcome the aforementioned difficulties. Reitmayr and Drummond [10] propose combination of inertial sensors and vision-based point and edge trackers for a handheld outdoor augmented reality. In their approach, inertial sensors provide a first estimation of the camera pose. A textured building model is then rendered according to the estimated camera pose, and an edge tracking algorithm determines the exact pose by matching edges in the video image with edges in the textured model. Ribo et al. [11] and Jiang et al. [5]

likewise use hybrid tracking systems of sensors and visual tracking techniques for robust outdoor augmented reality applications. For this study, focus was set on purely image based techniques only. Ferrari et al [2] propose tracking of parallelogram shaped or elliptic planar features, de-scribed in an invariant way. Their work shows interesting results, but lack generality, as suitable planes are never present in a high number, making this concept prone to occlusions, or could be missing completely in a given scene. For a virtual reconstruction of antique Greek monuments shown in an AR system, Vlahakis et al [14] use a high number of keyframes to keep differences between the current frame and the most similar keyframe low. This allows the use of simple and fast matching techniques, but effectively restricts the movement of the user to a few selected standpoints, as not every possible perspective could be anticipated and taken as key-frame in advance.

Gordon and Lowe [3] use Lowe's SIFT detector and descriptor [8] for an augmented reality application. SIFT stands for Scale Invariant Feature Transform, and is an invariant point detector and descriptor. In a first step, Gordon and Lowe take a set of reference images to create a sparse representation of the object or scene to be recognized and tracked. SIFT points are then extracted from each image, and matched against each other. During the augmented reality application, SIFT points are extracted in each frame and matched against the points of the point cloud. This establishes 2D-3D correspondences, which are used to calculate the camera position and pose. This approach has two drawbacks: The extraction of SIFT features is computationally demanding, which restricted the frame rate to 4 frames per second. Secondly, as the camera pose and position are calculated for each frame individually, the resulting virtual overlay jitters against the real world background. The method was documented for a small indoor scene and a very restricted outdoor scene comprising only one building front. Its performance on a larger scale is therefore unknown.

Vacchetti et al. [13] combine relative orientation between frames and absolute orientation towards a very low number of keyframes to reduce both drift and jitter. They choose the Harris interest point detector (Harris and Stephens, 1988 [4]) and image patches as their descriptor to match both between subsequent video frames and between video frames and keyframes.

3 Our Proposition

Our work was inspired by both the work of Gordon and Lowe [3] and of Vacchetti et al. [13]. For its simplicity, a similar approach as in Gordon and Lowe was chosen, that uses SIFT keypoints as means to establish correspondences between 3D and 2D points. Other invariant point descriptors exist, which are computationally lighter, including PCA-SIFT (Principal component analysis SIFT) [6], SURF (Speeded up robust features) [1], and GLOH (Gradient location-orientation histogram) [9], some of which are faster in computation. As the SIFT detector has proven its superior matching performance in a comparative study of Mikolajczyk and Schmid, [9], it was also chosen for this study. For more references on SIFT interest points and associated signature, see [8].

The aim of this study was to set up a prototype for an outdoor markerless augmented reality, based on the SIFT keypoint detector and descriptor. The application

outlined in section 1 of this paper requires a real-time operation on a mobile device. This was beyond the scope of this study. Instead, an implementation on a standard PC was realized that does not operate at real-time speed. It does however respect the requirements of a live application in that sense that only for each frame, only the information of previous frames or offline information was used.

The work for this study comprises 3 components: the calibration of the mobile device camera ; a 3D façade model of a small urban area and a 3D point cloud of SIFT points as the reference data ; a matching of video frames to the reference data and deduction of the camera location.

3.1 Camera Calibration

The camera of the mobile device has to be calibrated to know precisely its characteristics. The procedure is very simple. The user has to photograph a chessboard like calibration pattern from a number of different viewpoints. He then marks the outermost corners of the calibration pattern manually, and the calibration procedure automatically finds the different parameters of the camera, like focal length and lenses distortions.

3.2 3D Reference Model

To create the reference set of SIFT points, images of the building façades were taken. Each image was oriented absolutely with respect to the 3D model by manually providing control points. From these 3D -2D correspondences, the camera position and pose was calculated. In the next step, SIFT points were extracted from the image, and the rays from the centre of projection through the SIFT features in the image were constructed. Finally, the world coordinates of the points were obtained by intersecting the rays with the building façades. At early stages of the project, a manual region of interest was defined for each image to make sure that only points correctly originating from the façades are projected on the façade. The example on figure 1 presents the projection of SIFT points on the model.

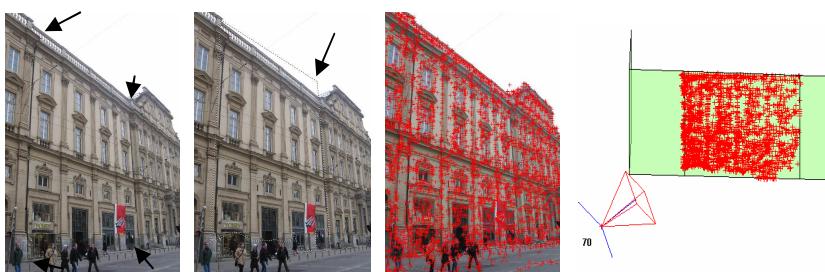


Fig. 1. Production of the reference set of SIFT points: Manual input of control points (arrows, left) – Manual definition of region-of-interest (center left) – extraction of SIFT features (center right) – All points within the region of interest are projected on the surface of the 3D model (right)

3.3 Matching of Video Frames to the Reference Data

During this study, four markerless augmented reality algorithms based on Lowe's SIFT [8] point detector and descriptor were implemented. All four algorithms share the same principle: Point correspondences between features in the current frame and a point cloud of reference points are used to calculate the pose of the camera in a robust fashion. The reference set was produced by projecting SIFT point features from a series of reference images onto the surface of a simplified 3D building model (see figure 2).

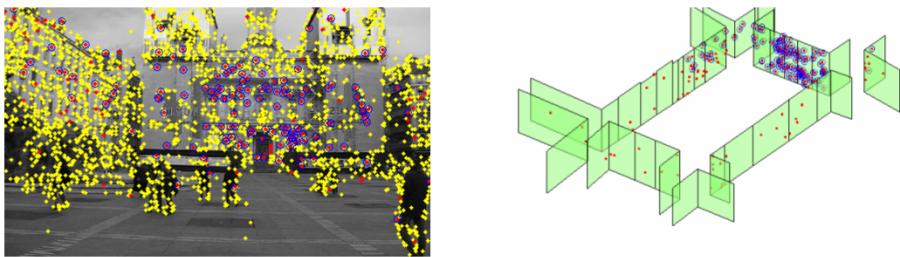


Fig. 2. Extraction of SIFT points (left) and matching to the 3D model (right). Only the points that match in sufficient number on the same area are kept. The other ones are considered as outliers.

The pose was calculated by minimizing the reprojection error of world points into the image in a least-squares sense within a RANSAC loop. This algorithm requires starting values for the camera's position and attitude. In this algorithm, a planar homography between four image points and four coplanar world points is used to initiate the pose estimation. This method has shown to deliver correct results reliably, provided the scene contains planar objects, a justified assumption for a building recognition application. The aforementioned methods to derive an initial pose were only used for the first frame. In the remaining video images, the pose of the previous frame was taken as the starting value, provided it was based on at least 15 point matches. This threshold was chosen arbitrarily and not tested in detail.

The camera pose was calculated in RANSAC loop. This algorithm is controlled by two values, namely the inlier threshold which refers to the maximum deviation a world point projected into the image plane may show from the true image point position to be regarded as an inlier point, and the number of iterations or camera poses that are calculated and tested. Extensive testing has shown that the shatter of the calculated position decreases with increasing iteration numbers, but remains almost constant if the number of iterations exceeds 250. Therefore, the value of 250 iterations is proposed to be used. Alternatively, an assumed ratio of correct points over all matched points of 25 % may be used.

Another method is implemented to reduce jitter. In this algorithm, a second set of points is detected in each image using the FAST detector [12]. These points are projected onto the 3D model. In the next frame, these points are found again by matching between the neighbouring frames using an image patch as a simple descriptor. The resulting additional 3D-2D correspondences are then used to increase the number of

point matches. This method effectively reduces jitter of the virtual image content with respect to the real objects, as shown by standard deviations of the calculated camera positions dropping to only a few decimetres in some sequences of the test video. Although the FAST detector, the projection and the matching method used here are simple calculations, they make only sense if they replace the costly SIFT calculation and matching, instead of being used additionally to SIFT.

4 Some Results

In the videos produced with the settings described above, the main reason for a complete loss of orientation was motion blur (see figure 3) and even in that case the algorithm recovers as soon as clear images are available again.

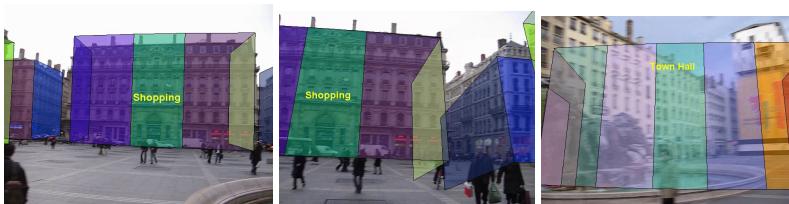


Fig. 3. Some images extracted from a computed augmented reality video. On these samples, we superimpose the 3D model to the real images to judge the accuracy of the matching. Correct matching (left) Wrong point of view estimation (center) Wrong matching due to motion blur (right).

Although in many of the frames produced for this study the virtual content jitters strongly, the algorithm has always recognized the buildings in the image correctly. This shows that the SIFT point detector and descriptor has the potential to be used for markerless augmented reality, provided its calculation can be accelerated. Errors in the reference dataset have been the most important reason for a bad fit of virtual content and filmed buildings, and not missmatches of the SIFT algorithm itself. This holds even for the uniform building façades that contain a lot of repeating structures. Although the reference images had a size of only 300 by 400 pixels, an average of 134 matches were found. The low ratio of inliers among these matches of approximately 25% is probably also caused by the low quality of the reference dataset. The 3D models used here contained important errors like wrong building heights and each reference image was oriented individually, mostly with only four control points. A better approach to produce the reference dataset would have been to use a more detailed 3D model, and to make a bundle adjustment over all input images to produce a consistent dataset.

5 Conclusion and Further Researches

While this research has shown that SIFT features are well suited for an augmented reality application under the given conditions, lower performance due to blur in the

frames coming from fast displacement of the camera, or complete different point of view or zoom on details, can be compensated by a bigger set of reference images, or by altering the reference images synthetically. This raises the question: How must the reference set be made? The number of reference images depends on the scale of the objects to be contained in the reference set, which again depends on the application of the augmented reality application. If information about buildings is to be displayed, it is sufficient to cover the buildings with images as done in this study. However, if smaller objects like, for example, building details are to be detected, the number of needed reference images increases. An increased number of reference images has unfavourable consequences: The effort to produce the reference set is increased, the size of the reference set gets bigger, which makes storage and transfer of it more difficult, and the matching process takes more time when the search domain increases.

Similar question applies for the lighting conditions. Is this dataset still sufficient in dawn or at night time? Once the requirements on the reference set are known in more detail, automatic methods to derive the reference dataset would be of great advantage.

Finally, further investigations are necessary to transform the augmented reality algorithms into an intuitive tool that helps the user to fulfil his information demand in a simple and easy-to-use fashion.

References

- [1] Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded Up Robust Features. In: Proceedings of the ninth European Conference on Computer Vision, May, pp. 404–417 (2006)
- [2] Ferrari, V., Tuytelaars, T., Van Gool, L.: Markerless augmented reality with a real-time affine region tracker. In: Proceedings of the IEEE and ACM International Symposium on Augmented Reality, pp. 87–96 (2001)
- [3] Gordon, I., Lowe, D.: Scene modelling, recognition and tracking with invariant image features. In: Proc. 3rd IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2004), Arlington, pp. 110–119 (2004)
- [4] Harris, C., Stephens, M.: A combined corner and edge detector. In: Proc. of The Fourth Alvey Vision Conference, 1988, Manchester, UK, pp. 147–151 (1998)
- [5] Jiang, B., Neumann, U., You, S.: A Robust Hybrid Tracking System for Outdoor Augmented Reality. In: Proc. of the IEEE Virtual Reality Conference 2004 (VR 2004), pp. 3–10 (2004)
- [6] Ke, Y., Sukthankar, R.: PCA-SIFT: A more distinctive representation for local image descriptors. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 506–513 (2004)
- [7] Lim, J.-H., Chevallat, J.-P., Merah, S.N.: SnapToTell: Ubiquitous Information Access from Camera. In: A Picture-Driven Tourist Information Directory Service in Mobile & Ubiquitous Information Access (MUIA 2004) Workshop as part of the conference Mobile Human Computer Interaction with Mobile Devices and Services (Mobile HCI 2004), Glasgow University of Strathclyde, Scotland, September 2004, pp. 21–27 (2004)
- [8] Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–100 (2004)
- [9] Mikolajczyk, K., Schmid, C.A.: Performance evaluation of local descriptors. In: Proc. of IEEE Computer Vision and Pattern Recognition, vol. 2, pp. 257–263 (2003)

- [10] Reitmayr, G., Drummond, T.: Going Out: Robust Model-based Tracking for Outdoor Augmented Reality. In: Proc. IEEE ISMAR 2006, Santa Barbara, USA, pp. 109–118 (2006)
- [11] Ribo, M., Lang, P., Ganster, H., Brandner, M., Stock, C., Pinz, A.: Hybrid tracking for outdoor augmented reality applications. *IEEE Comp. Graph. Appl.* 22(6), 54–63 (2002)
- [12] Rosten, E., Drummond, T.: Fusing points and lines for high performance tracking. In: Proc. 10th IEEE International Conference on Computer Vision (ICCV 2005), Beijing, vol. 2, pp. 1508–1515 (2005)
- [13] Vacchetti, L., Lepetit, V., Fua, P.: Combining edge and texture information for real-time accurate 3D camera tracking. In: Proc. 3rd IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2004), Arlington, VA, pp. 48–57 (2004)
- [14] Vlahakis, V., Ioannidis, N., Karigiannis, J., Tsotros, M., Gounaris, M., Stricker, D., Gleue, T., Daehne, P., Almeida, L.: Archeoguide: An Augmented Reality Guide for Archaeological Site. *IEEE Computer Graphics and Applications* 22(5), 52–60 (2002)

Off-Line Arabic Handwritten Characters Recognition Based on a Hidden Markov Models

M. Amrouch, M. Elyassa, A. Rachidi, and D. Mammass

IRF-SIC Laboratory

Faculty of Sciences, Agadir , Morocco

amrouch_mustapha@yahoo.fr, melyass@gmail.com,
rachidi@caramail.com, driss_mammass@yahoo.fr

Abstract We present a system based on Hidden Markov Models (HMMs) for off-line isolated Arabic handwritten characters recognition. The Hough accumulator of the character image is partitioned into equal horizontal bands that will be used to extract directional information. This information is translated into sequences of observations that are used to train the model for each character during the learning step.

Keywords: Isolated Arabic hand-written characters; Hough Transformation; Hidden Markov Models.

Résumé

Nous présentons dans ce travail un système à base de Modèles de Markov cachée (HMMs) pour le problème de la reconnaissance hors-ligne des lettres arabes isolées manuscrites. L'accumulateur de Hough de l'image du caractère est partitionnée en bandes horizontales égales que nous utiliserons pour extraire une information directionnelle. L'information obtenue est traduite en séquences d'observations qui sont utilisées pour entraîner le modèle de chaque caractère lors de la phase d'apprentissage.

Mots clés

Caractères arabes isolés manuscrits, Transformation de Hough, Modèles de Markov Cachés.

1 Introduction

The morphology of the Arabic writing presents some characteristics which are the source of their treatments complexity. The Arabic writing is semi cursive script in its printed and handwritten forms. The characters of an Arabic word (or pseudo-word) are horizontally or vertically ligatured which darken the process of segmentation in characters. The forms of letters change according to their positions in the word. Besides, more than half of the Arabic characters include diacritic points in their shape [4] [5].

The Hidden Markov Models (HMMs) are among the most used methods in recognition ([1] [2] [3] [11] [12]). We present in this paper a Hidden Markov Models for off-line isolated Arabic handwritten characters recognition (see figure 1) which is distinguishes by the exploitation of the rates of presence of the dominant orientations of the characters Hough transform.

١	٢	٣	٤	٥	٦	٧	٨	٩
٠١	٠٢	٠٣	٠٤	٠٥	٠٦	٠٧	٠٨	٠٩
١٠	١١	١٢	١٣	١٤	١٥	١٦	١٧	١٨
٢٠	٢١	٢٢	٢٣	٢٤	٢٥	٢٦	٢٧	٢٨

Fig. 1. Arab characters

In the section 2, we present the architecture of our system, the data base of Arabic handwritten characters and the pre-processing. The section 3 is reserved to the characteristics extraction and to the choice of the models, obtained from Markovian observations, which will represent the characters. The section 4 is dedicated to the learning and classification steps and the section 5 to the numerical results.

1.1 Hidden Markov Models (HMMs)

A Hidden Markov Model is a double stochastic process with an underlying process which is not observable. This last one is only observed through another set of stochastic processes which produces the sequence of the symbols. An HMM is characterized by the state transition probabilities matrix A , the symbols probabilities of observation B and the vector of probability of the initial state Π .

N : The number of states;

T : The number of observations or possible symbols;

q_t : The state of the system at the time t ;

M : Size of the observed sequence;

$A = \{a_{ij} = p(s_j / s_i)\}; \sum_{j=1}^N a_{ij} = 1$ is the matrix of the probability of transitions.

$\Pi = \{\pi_i = p(s_i)\}; \sum_{i=1}^N \pi_i = 1$ is the vector of the initial probabilities.

$B = \{b_i(o_k) = p(o_k / s_i)\}; \sum_{k=1}^T b_i(o_k) = 1$ are the probabilities of the observation symbols.

A Hidden Markov Model is characterized by a matrix of transition A , a matrix of observation B and a vector of initialization Π .

We note: $\lambda = \{\Pi, A, B\}$

The use of the HMMs in the recognition is conditioned by the resolution of three problems:

- Recognition: Let an HMM $\lambda = \{\Pi, A, B\}$ and an observed sequence $O = \{o_1, o_2, \dots, o_n\}$, what is the value of the probability $p(O/\lambda)$ so that the model λ generates O ?
- Analysis an decoding : For an HMM and an observed sequence O , what is the sequence of the states which has the maximal probability to generated O ?
- Learning: From a chain of observations $O = \{o_1, o_2, \dots, o_n\}$, how to adjust the parameters of the HMM $\lambda = \{\Pi, A, B\}$ to maximize $p(O/\lambda)$?

The solutions of these problems are well retailed in the literature [6] [7] [8] [9].

2 System of Recognition

2.1 Architecture

We begin with a pre processing step in order to remove noises due to the acquisition process of the characters (see figure 2). Directional primitives are extracted from the processed images by using the Hough transformation and then sequences of observations are generated. In the training step, we involve the HMM of each

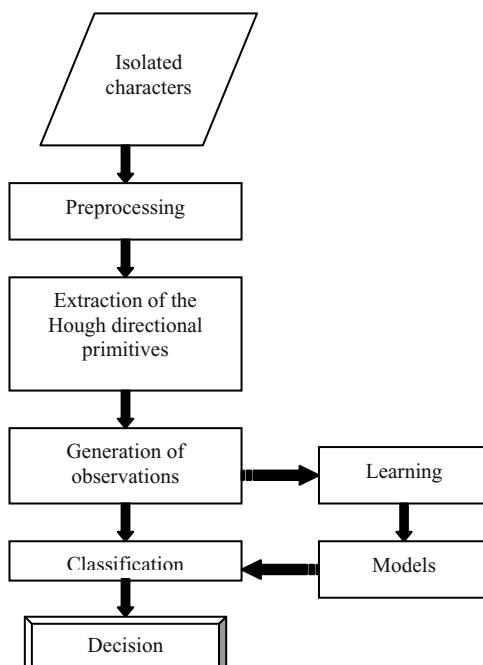


Fig. 2. System of recognition

character, with a Baum-Welch process, by the various sequences of observations. The classification is done by the search for the characters discriminating models and we decide for the class of the character.

2.2 Data Base

The system is tested on a data base of 6188 isolated Arabic handwritten characters (26 persons giving, each one 14 times, 17 Arabic characters). The images of characters are in Gray level of dimensions 96 X 96 pixels. Some operators are used to reduce the noise and there is no operation of segmentation because we have only isolated forms.

3 Extraction of Primitives

The direct identification of the character from its image (matrix of pixels) seems very difficult because of the Arabic characters morphology and the variability in the used writing style. So, for each images, the Hough Transform [10] give a set of discriminating characteristics, from the character representation in pixels, that allows recognizing it easily. The algorithm of extraction is:

- Initialization of the table of the Hough Transform accumulator
- For each black pixel of the image $(x_i, y_i) : 0 \leq x_i \leq n$ et $0 \leq y_i \leq m$ with n width image and m its height.
- $0 \leq \theta < 180^\circ$
- $\rho_k = x_i \cos \theta_k + y_i \sin \theta_k$
- To increment the cell of the table of accumulator corresponding to the couple (ρ_k, θ_k)

The result of the algorithm on some characters is shown in figure 3.



Fig. 3. Hough Transform on the characters KAF and AIN

In the experimentation, the images of characters are of fixed size: 96×96 pixels and the displacement $\Delta\theta$ is equal to 30° , consequently the Hough accumulator will contain information on the six orientations ($0^\circ, 30^\circ, 60^\circ, 90^\circ, 120^\circ$ et 150°).

The images are cut out in 16 horizontal and 16 vertical bands and the intersections of these bands form 256 areas of dimension 6×6 pixels each one. For each area, we calculate the rates of the six orientations presence and we determine the dominant direction, and finally we memorize this information which will be used for the generation of observations.

3.1 Generation of the Directional Observations Sequences

For the generation of the sequence of directional observations task, we select the minimum of the primitives representing the dominant directions. In each area, consider the rate of the dominant direction. Thereafter, we represent, by a symbol in the sequence of observations, a set of adjacent areas considered of the right-hand side towards the left and of the same rates. The symbols which can be emitted are 6 corresponding to the 6 basically orientations (see figure 4).

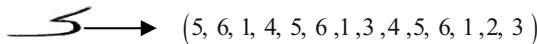


Fig. 4. Sequence of the character KAF

3.2 Choice of Models

The images are translated into sequences of observations which give the Hidden Markov Models representing every character. In the literature, various models and architectures were proposed for the use of the HMM in the handwritten writing recognition ([11][12]). In our work, we opted for a right-left topology where the number of states is strictly lower than the number of horizontal bands of the character. Indeed, the bands of the extremities supply no observation.

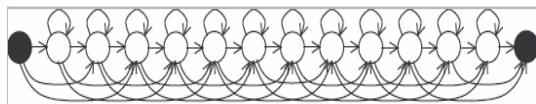


Fig. 5. Model of character with N states [3]

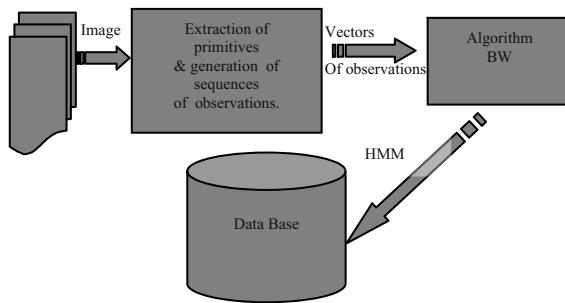
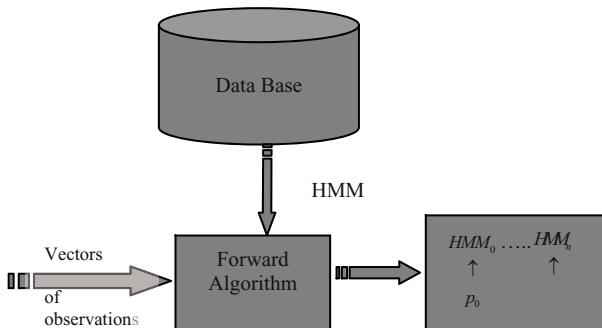
4 Learning and Classification

4.1 Learning

We train the hidden Markov models of various characters by the classical procedure of Baum-Welch to adjust their parameters. Every character possesses its own model. Consequently, the algorithms look for, in the space of the HMM modelling of each character, the model which has the maximum probability to generate the sequence of observations "generation of the sequence of observation ". The best HMM is selected to form a base of learning (see figure 5).

4.2 Classification

In the learning step, we obtain so many HMMs as there were of characters. The classification is made at first by the search of the discriminating model among every

**Fig. 6.** Scheme of the learning data**Fig. 7.** Scheme of the classification process

selected HMMs of each character. We calculate, by the Forward algorithm, the probability such that the models can generate the sequence of observations of the character to be recognized. Afterward, we have a set of models where every model is associated with a score. The elected model is the one possessing the biggest score (see figure 6).

5 Numerical Results

We used 42 Arabic handwritten isolated characters (14 KAF, 14 HA, 14 AIN) extracted from the base of characters. 50 % served for the phase of learning, and 50 % for the tests. Encouraging results appear with a rate of 85.71 % of good recognition. The best score was 100 % for the character (Ha) (see figure 8).

The errors of recognition are big for the letter AIN, what is explained by the incapacity of the used characteristics to describe every character (very big variability in the styles of the writing of this character). A good estimation of these data can decrease the rate of error of our system.

Arabic character	Translation in French	Rate of recognition
	Ha	100 %
	KAF	85,71 %
	AIN	71,42 %

Fig. 8. Example of rate recognition

5.1 Conclusion

In this paper, we approached the problem of the off-line isolated Arabic handwritten characters recognition based on Hidden Markov Models (HMMs) and directional information from the Hough transformation of every character.

The results are promising and show that the Hidden Markov Models are well adapted to the variation in the length of the handwritten writing. However, its capacity of discrimination is not very high because every HMM uses the data of learning of a single character. One of the weaknesses of the HMMs, results from the level of the probability estimation of the emission of observations. To ovoid these problems, we think to use an hybrid method combining the HMMs and the neural networks.

References

- [1] Belaïd, A., Saon, G.: Utilisation des processus markoviens en reconnaissance de l'écriture. *Traitemet de signal* 14(2), 161–178 (1997)
- [2] El Yacoubi, A.: Modélisation markovienne de l'écriture manuscrite, application à la reconnaissance des adresses postales. *Thèse de doctorat*, Université de Rennes 1 (1996)
- [3] poisson, E.: Architecture et Apprentissage d'un système hybride de Neuro-Markovien pour la reconnaissance de l'écriture Manuscrite en-ligne, *Thèse de doctorat*, université de Nantes (2005)
- [4] Ben Amara, N., Belaïd, A., Ellouze, N.: Modélisation pseudo bidimensionnelle pour la reconnaissance de chaînes de caractères arabes imprimés. In: *Proc. 1er Colloque international francophone sur l'écrit et le document (CIFED 1998)*, Québec, Canada, pp. 131–140 (1998)
- [5] Miled, H. : Stratégies de reconnaissance de l'écriture semi cursive : application aux mots manuscrits arabes. *Thèse de doctorat*, Université de Rouen (1998)
- [6] Cornuéjols, A., Miclet Eyrolles, L.: Apprentissage Artificiel : Méthodes et Algorithmes' IRISA, ENSSAT (2002)
- [7] Dupré, X.: Contributions à la reconnaissance de l'écriture cursive à l'aide de modèles de Markov cachés, *Thèse de doctorat*, Université René Descartes - Paris V (2004)
- [8] Korilis, Y., St-Jean, C., DeBarr, D., Carpenter, B., Chu-Carroll, J.: Modèles de Markov Cachés
- [9] Agen, F., Michot, J.: Chaînes de Markov cachées Algorithme de Baum-Welch, Ecole polytechnique université de Tours (2004-2007)

- [10] Maitre, H.: Un panorama de la transformation de Hough. In: Traitement de signal, vol. 2 (1985)
- [11] Mahjoub, M.A.: Application des modèles de Markov Cachés stationnaires et non stationnaires et non stationnaires à la reconnaissance en ligne de l'écriture arabe, Thèse de doctorat ENIT (1999)
- [12] Touj, S., Touj, S., Ben Amara, N., Amiri, H.: Reconnaissance Hors Ligne de Caractères Arabes Isolés Manuscrits, Institut Supérieure des Etudes Technologiques de Gabès, Ecole Nationale d'Ingénieurs de Monastir, Ecole Nationale d'Ingénieurs de Tunis, CIFED 2002 (2002)

Watermarking Based on the Density Coefficients of Faber-Schauder Wavelets

Mohamed El hajji¹, Hassin Douzi¹, and Rachid Harba²

¹ RF-SIC, Faculté des sciences Ibn Zohr, Agadir, Maroc

² PRISME Institut, Polytech'Orléans, 12 rue de Blois 45067 Orleans CEDEX 2

Abstract. The field of watermarking is of great development in recent years. There is still relevance to find high-performance methods that link robustness and discretion. In this paper, we will present a method of watermarking based on a mixed scale wavelet transformations of Faber-Schauder and the selection of high-density regions of the dominant coefficients. Simulation results demonstrate the robustness of our image watermarking procedure under various attacks and could serve for the image watermarking of identity and access cards.

Keywords: Digital watermarking, Faber-Schauder wavelet transformation, dominant coefficients.

1 Introduction

The digital watermarking tends to exploit the imperfections of the human visual system to embed invisible identification information into a digital content (video, audio or image), in order to protect it [10]. In the original document, we insert a signature (Watermark), which is formed by a sequence of bits. The insertion must respect the imperceptibility compared to the human visual system and the robustness (i.e. difficult to extract or to delete).

The wavelet transformation (DWT) is now widely used in image processing [12]. They are distinguished by a multi-scales decomposition of the image to separate the structures of different sizes. Most of the watermarking methods in the wavelet domain are working with scales separated [15]. For each scale, pixels are decorrelated by this transform and information is only present in a small number of wavelet coefficients: the dominant coefficients [6.7]. These coefficients are located mainly around the edge of the image and characterize all the information of the image. This is especially true for identities images which are the topic of our interest (the images scanned from sensitive documents: passports, identity cards, access cards....). The originality of our method is the use of mixed-scale visualization [6] ,that permit to distinguish very particular regions around the edges of the image, where we have a high density of dominant wavelet coefficients. These regions are characterized by the fact that they are highly textured regions, therefore ideal to hide information. It is also the regions with a high density of the dominant coefficients, therefore unavoidable in the characterization of the image and a priori stable for any transformation of the image. Another originality of our method is the use of the density of dominant coefficients as a criterion at the insertion and the extraction.

In this paper, we present a method based on the density of wavelet coefficients. The Faber-Schauder wavelets are chosen since they are particularly well adapted for the watermarking [7].

After describing the mixed-scale visualization, we will propose a strategy for watermarking and test its robustness following various attacks.

2 Faber-Schauder Wavelet Transform

The algorithm of Faber-Schauder wavelet transform (FSWT) is simple to express by a lifting scheme [6, 16]. We can consider a image as a sequence $f^0 = (f_{m,n}^0)_{m,n \in \mathbb{Z}}$ of $L^2(\mathbb{Z}^2)$, Transform FSWT is done in three steps as shown in the following figure:

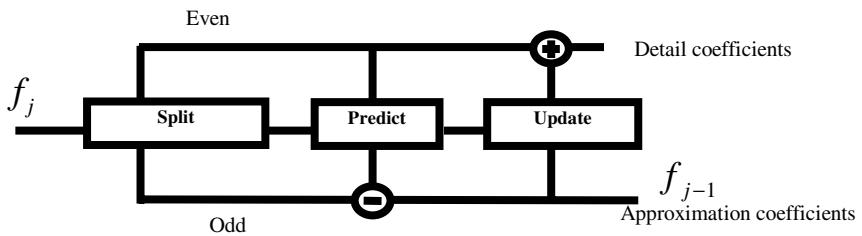


Fig. 1. The transformation by lifting scheme

The Scheme lifting of the FSWT [6] is given by the following algorithm:

$$\begin{aligned}
 & \left\{ \begin{array}{l} f_{ij}^0 = f_{ij} \quad \text{pour } i, j \in \mathbb{Z} \\ \text{pour } 1 \leq k \leq N \text{ et } i, j \in \mathbb{Z} \\ f_{ij}^k = f_{2i,2j}^{k-1} \\ g_{ij}^{k1} = (g_{ij}^{k1}, g_{ij}^{k2}, g_{ij}^{k3}) \\ g_{ij}^{k1} = f_{2i-1,2j}^{k-1} - \frac{1}{2}(f_{2i,2j}^{k-1} + f_{2i+2,2j}^{k-1}) \\ g_{ij}^{k2} = f_{2i,2j+1}^{k-1} - \frac{1}{2}(f_{2i,2j}^{k-1} + f_{2i+2,2j+2}^{k-1}) \\ g_{ij}^{k3} = f_{2i+1,2j+1}^{k-1} - \frac{1}{4}(f_{2i,2j}^{k-1} + f_{2i,2j+2}^{k-1} + f_{2i+2,2j}^{k-1} + f_{2i+2,2j+2}^{k-1}) \end{array} \right. \\
 & \text{TOS :}
 \end{aligned} \tag{1}$$

3 The Watermarking by the Density of the Coefficients of Faber-Schauder Wavelet

3.1 Watermark Embedding

The basic idea consists in inserting a mark on the edges detected by the density by block of coefficients of FSWT, while taking into account two important aspects: the robustness and invisibility.

The algorithm for watermark embedding can be presented as following:

- Firstly we calculate FSWT of original image \mathbf{I} , to get a transformed image \mathbf{F} in mixed scales representation [6].
- The second step consists of selecting the high density blocks of wavelet coefficients followed by a classification of these blocks in ascending order, and so to insert the mark in the important blocks (in the sense of density). The embedding can also be effected on all blocks selected. this way, the mark to embed may be redundant if it has a length less than the total number of blocks extracted.
- The third step consists to applying a spread spectrum method into the watermark \mathbf{W} by using a random sequence broadband \mathbf{U} [8], so, the gotten message reinforces the resistance of the signature to perturbations provoked on the image. Each block is going to carry information in related with a given bit of message obtained in the form of an additive term. The watermark embedding follows the following equation:

$$C_f^*(i, j) = C_f(i, j)(1 + \alpha W_k(i, j)) \quad (2)$$

Here the constant α gives the watermark strength.

$C_f^*(i, j)$: The marked coefficient (with $i, j=1..$ the block size).

$C_f(i, j)$: Initial coefficient of the selected block.

$W_k(i, j)$: the sequence obtained by the spread spectrum method.

After the marking of a block, we replace it at the corresponding block in the image \mathbf{F} .

- Finally, we reconstruct the changed image by using the inverse wavelet transform.

The selected blocks are those containing large densities of dominant wavelet coefficients as shown in Fig. 2.

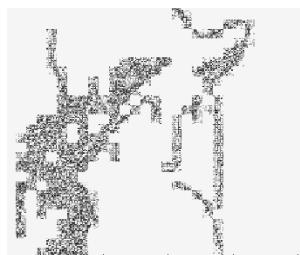


Fig. 2. The selection of blocks of high-density of the dominant coefficients (For Lena image: 8x8 blocks, SC = 10, SD = 20%)

The selection of blocks is based on the following selection criterion:

- A threshold SC is fixed to select the dominant coefficients: A coefficient is considered dominant if its absolute value is superior to SC .
- A threshold SD of density is fixed to select insertion blocks: All block whose density passes this threshold will be selected as a dominant block.

3.2 Watermark Detection

The insertion of the watermark is made by a spread spectrum in the dominant blocks. At the reception of a marked image, we calculate the correlation with the random sequence \mathbf{U} used for the modulation [8].

More precisely we proceed as follows:

- Calculate Faber-Schauder wavelet transformation of the supposed marked image.
- Select the dominant blocks: blocks with high density of wavelet dominant coefficients.
- Calculate the correlation between the selected blocks and the random sequence \mathbf{U} :

$$r(\mathbf{B}, \mathbf{U}) = \frac{\sum_m \sum_n (B_{mn} - \bar{B})(U_{mn} - \bar{U})}{\sqrt{\left(\sum_m \sum_n (B_{mn} - \bar{B})^2 \right) \left(\sum_m \sum_n (U_{mn} - \bar{U})^2 \right)}} \quad (3)$$

With \mathbf{B} the constructed vector from a given block.

- Choose the maximum of the calculated correlations, noted r_{\max} .
- If r_{\max} superior to a threshold T then the image is considered marked.

4 Results and Conclusion

4.1 Visual's Deterioration

The first tests done to measure the image deterioration (by the watermark) are the calculation of the PSNR (table1). We can notice that the deteriorations are very little visible for the marked image (lena 512) (Fig. 7). As long as the threshold is lower than 10, PSNR remains well above 30 dB.

The performances of a marking algorithm are measured by the degree of resistance to classic attacks and to filtering, We will use for the proposed method the bench attacks CheckMark[13], The test bench is generated from one or several marked images, Then we will apply the algorithm of detection on each attacked image, one tried to detect the mark and we look if the detection succeeded or no,

This decision is made while defining a threshold of detection; the threshold is equal to **0.40**. It is an empirically chosen from several tests.

Table 1. PSNR for marked images

The watermark strength.	PNSR (db)	correlation
2	37.4035	0.437552
7	32.8847	0.633063
10	30.2074	0.743258
15	27.3072	0.805537
20	25.0252	0.809627
30	21.7905	0.929487

4.2 Robust Against Compression

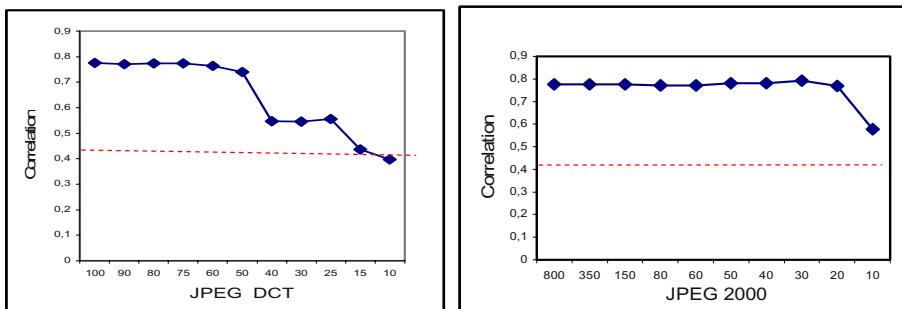


Fig. 3. The correlation for different qualities of compression by JPEG DCT: The resistance to JPEG 2000 compression algorithm: it's the JPEG compression (DCT) is relatively important (more far from the predefined threshold)

4.3 Filtering

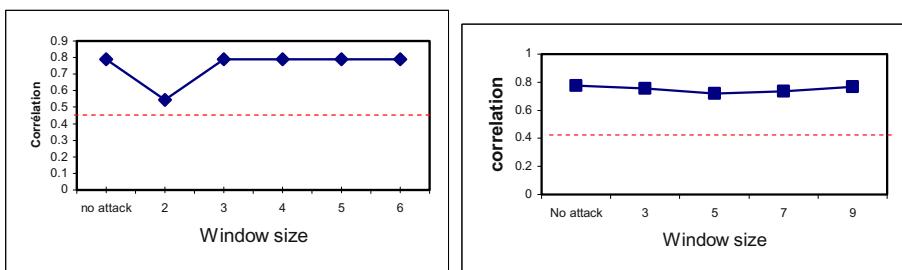
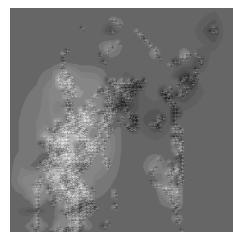


Fig. 5. The correlation obtained after an attack with the Gaussian filtering: the correlation still high than the threshold

Fig. 6. The correlation obtained after an attack by the Wiener filtering: the correlation still very important



threshold =20, PNSR= 25.02 db



The difference image between the marked image and the original image

Fig. 7. If the threshold of marking is too important, the distortions of the marked image are very important

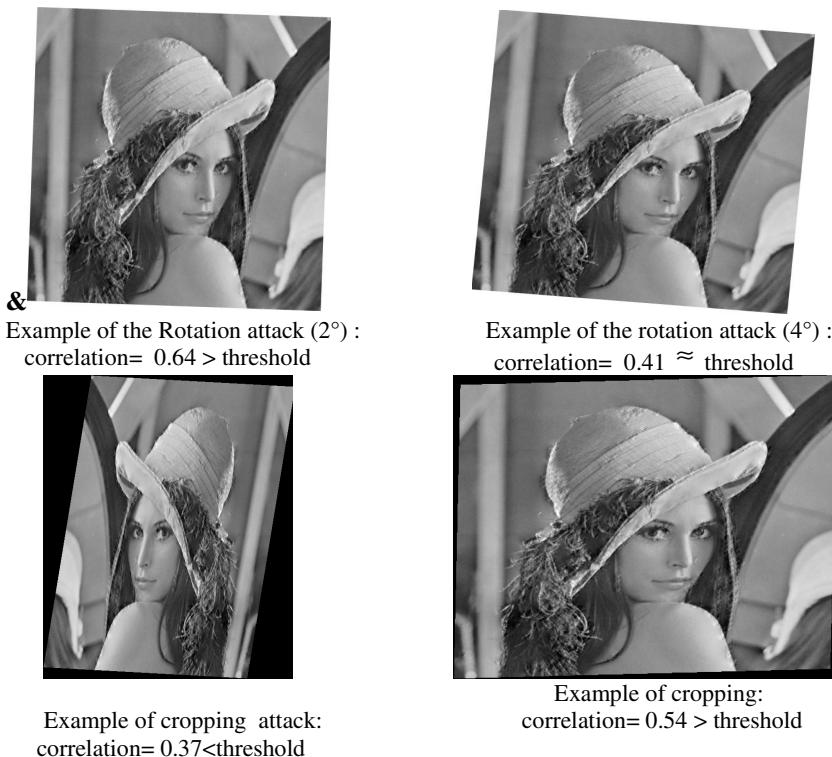


Fig. 8. Example of geometric attack



Fig. 9. Example of Stirmark attack: correlation= 0.42

4.4 Geometric Distortions

The geometric attacks such as cropping, resizing and rotation attacks knock down the correlation scores significantly (cf. Fig 8 and 9). It is therefore necessary to improve the performances of this method, particularly by finding a solution to the problem of synchronization: Indeed the geometric distortion introduced a spatial desynchronization

changing the density of blocks dominant of wavelet coefficients. So when we turn the image by 10° the detection of the signature by the proposed method become ineffective. But for relatively small distortions (for example rotations lower to 4°) the detection of the signature still possible with our proposed method. The application that we aim has the particularity to provoke geometric distortions of very low amplitude [12] which shows that our approach could be successful.

5 Conclusion

In this papers we have presented a method of images watermarking based on mixed-scales visualization by Faber-schauder wavelet. The experimental results have confirmed that this method is robust against various attacks: JPEG compression and geometric of small angle or displacement. It could therefore be used for images watermarking in this case. This method will be tested on images of access systems with magnetic badge where we have a geometric attack with small amplitude.

References

1. Petitcolas, A.P., Anderson, R.J., Kuhn, M.G.: unformation hiding a survey. Proceedings of the IEEE 87, 1062–1078 (1999)
2. Cox, I., Miller, M.L.: A review of watermarking and the importance of perceptual modeling. In: Proceedings of SPIE, Humain visual and Electronic imaging II, vol. 3016, pp. 92–99 (1997)
3. Hartun, F., Kutter, M.: multimedia watermarking techniques. Proceedings of the IEEE 87, 1079–1107 (1999)
4. Cox, I., Killian, J., Leighton, F.T., Shamoon, T.: A secure robust watermark for multimedia. In: Anderson, R. (ed.) IH 1996. LNCS, vol. 1174, pp. 185–206. Springer, Heidelberg (1996)
5. Piva, A., Barni, M., Bartolini, F., Cappellini, V.: DCT-based watermark recovering without resorting to the uncorrupted original image. In: IEEE Signal Processing Society 1997 International Conference on Image Processing (ICIP 1997), Santa Barbara, California (1997)
6. Douzi, H., Mammass, D., Nouboud, F.: Faber-Schauder wavelet transformation application to edge detection and image characterization. Journal of Mathematical Imaging and Vision 14(2), 91–102 (2001)
7. Douzi, H., Mammass, D., Nouboud, F.: Amélioration de la compression par la transformation multiéchelles de Faber-Schauder. In: Vision Interface 1999 proceeding, 18-21 May, Trois-Rivières, Québec, Canada, pp. 424–431 (1999)
8. Cox, I.J., Kilian, J., Leighton, T., Shamoon, T.: Secure Spread Spectrum Watermarking for Multimedia. IEEE Trans. on Image Processing 6(12), 1673–1687 (1997)
9. Sweldens, W.: The lifting scheme: A construction of second generation wavelets. SIAM. J. Math. Anal. 29(2), 511–546 (1997)
10. Cox, I.J., Doërr, G., Furou, T.: Watermarking Is Not Cryptography. In: Shi, Y.Q., Jeon, B. (eds.) IWDW 2006. LNCS, vol. 4283, pp. 1–15. Springer, Heidelberg (2006)
11. Ros, F., Borla, J., Leclerc, F., Harba, R., Launay, N.: Watermarking for Plastic Card Supports. In: 9ème Conference Maghrébine sur Les Technologies de L’Information, MCSEAI, Agadir (2006)

12. Chang, C., Maleki, A., Girod, B.: Adaptive Wavelet Transform for Image Compression via Directional Quincunx Lifting. In: Proc. IEEE International Workshop on Multimedia Signal Processing, MMSP 2005, Shanghai, China (2005)
13. Pereira, S., Voloshynovskiy, S., Madueño, M., Marchand-Maillet, S., Pun, T.: Second generation benchmarking and application oriented evaluation. In: Information Hiding Workshop III, Pittsburgh, PA, USA (April 2001)
14. Sweldens, W.: The lifting scheme: A construction of second generation wavelets. SIAM Journal on Mathematical Analysis 29(2), 511–546 (1998)
15. Kundur, D., Hatzinakos, D.: Digital Watermarking Using Multiresolution Wavelet Decomposition. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, May 1998, vol. 5, pp. 2969–2972 (1998)

Arabic Handwritten Characters Classification Using Learning Vector Quantization Algorithm

Mohamed A. Ali

Computer Science Dept., Faculty of Science, Sebha University, Sebha, Libya
fadeel1@Sebhau.edu.ly

Abstract. In this module, Learning Vector Quantization LVQ neural network is first time introduced as a classifier for Arabic handwritten character. Classification has been performed in two different strategies, in first strategy, we use one classifier for all 53 Arabic Character Basic Shapes CBSs in training and testing phases, in second strategy we use three classifiers for three subsets of 53 Arabic CBSs, the three subsets of Arabic CBSs are; ascending CBSs, descending CBSs and embedded CBSs. Three training algorithms; OLVQ1, LVQ2 and LVQ3 were examined and OLVQ1 found as the best learning algorithm.

Keywords: Arabic handwritten recognition, Neural Network, Classification, Character Recognition.

1 Introduction

Arabic Off-line handwriting character recognition has been a difficult problem to machine learning. It is hard to mimic human classification where specific writing features are utilized [5]. Recent surveys have shown that present technologies have still a long way to catch up in terms of robustness and accuracy [1]. Compared with machine-printed character recognition, the prime difficulty in the research and development of handwritten character recognition systems is in the variety of shape deformations [3].

Classification module is one of stages in an Arabic optical character recognition system that we are developing. Classification stage came after preprocessing, segmentation and features extraction. The classification problem can be stated as finding functions which map feature vectors to classes. Preferably these functions should map clusters of same class objects in the feature space [6-7].

2 Classification Approaches

Classifier can be designed using number of possible approaches [4]. In practice, the choice of a classifier is a difficult problem and it is often based on which classifier(s) happen to be available, or best known to the user. Three different types of classification techniques have been identified.

The first type of pattern classification technique is based on the probabilistic approach. Optimal Bayes decision rule is one example of this approach. The second classification techniques constructs decision boundaries (geometric approach) directly by optimizing certain error criterion. The driving force of the training procedure is, however, the minimization of a criterion such as the apparent classification error or the mean squared error (MSE) between the classifier output and some preset target value. The third, simplest and the most spontaneous types are based on the concept of similarity by which patterns can be classified by minimum distance classifier using a few prototypes per class, some time called nearest neighbor (NN) classifier [2]. The choice of prototypes is crucial for the performance of these types of classifiers. Another advanced technique for computing prototypes is unsupervised Neural Network called "Self-Organizing Maps" (SOM), and its supervised version network "Learning Vector Quantization (LVQ) [1]. SOM is capable of detecting correlations and regularities in their input and adapt their upcoming responses to that input accordingly. The main aim of using SOM is to encode a large set of input vectors $\{x\}$ by finding a smaller set of representatives called "prototypes" or "codebook vectors" $\{w_I(x)\}$ that provide a good approximation to the original input space [2]. LVQ network has been chosen to be implemented as classifier for Arabic OCR system due to the aforementioned reasons.

2.1 Learning Vector Quantization

Learning Vector Quantization (LVQ) is a nearest-neighbor method operating essentially in the input domain [2]. It consists of a preset number of processing units, each unit having a d-element reference vector, and each unit being associated with one of the classes of the input samples.

Learning vector quantization (LVQ) is an algorithm for training competitive layers of SOM networks in a supervised manner. A competitive layer automatically learns to classify input vectors. However, the classes that found by the competitive layer are dependent only on the distance between input vectors. If two input vectors are very identical, the competitive layer most likely will put them in the same class. LVQ networks, on the other hand, learn to classify input vectors into target classes chosen by the user. The basic architecture of LVQ network is shown in Figure 1.

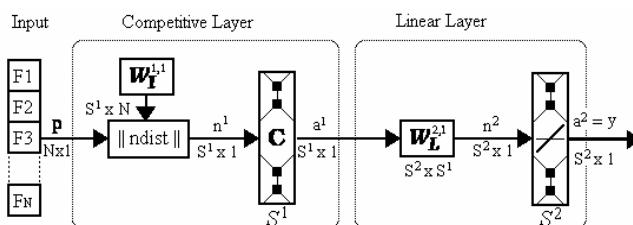


Fig. 1. Learning Vector Quantization Network

Where: S^1 and S^2 are numbers of competitive and linear neurons respectively.
 N : is number of elements in input vector P (features: $F1, F2 \dots$ and FN)

$W_I^{1,1}, W_L^{2,1}$ are competitive and linear weights respectively.
 $ndist$: is a function for calculating near-distance

2.2 Learning Vector Quantization Algorithms

Learning Vector Quantization utilizes three different learning techniques namely; LVQ1, LVQ2, and LVQ3. LVQ1 algorithm has an optimized version called OLVQ on which we will give brief description here, further details on other algorithms can be found in Kohonen [3]. OLVQ1 was implemented as a classifier in this research.

I. The LVQ1

In this algorithm various domains of the input vector x is approximated by placing a number of codebook vectors m_i (free parameter vectors) in the input space. The quantized values of codebook vectors are used for this approximation. Typically a number of codebook vectors are specified to each class of x values, and x is then decided to belong to the same class to which the nearest m_i belongs. Let the following equation define the nearest m_i to x ,

$$c = \arg \min_i \{ \|x - m_i\| \} \quad (1)$$

The following equations illustrate the mechanism of the basic LVQ1 process:

$$m_c(t+1) = m_c(t) + \alpha(t)[x(t) - m_c(t)] \quad (2)$$

if x and m_c belong to the same class,

$$m_c(t+1) = m_c(t) - \alpha(t)[x(t) - m_c(t)] \quad (3)$$

if x and m_c belong to different classes,

$$m_i(t+1) = m_i(t) \quad \text{for } i \neq c. \quad (4)$$

Where $\alpha(t)$ is individual learning rate. Here $0 < \alpha(t) < 1$, and $\alpha(t)$ may be constant or decrease monotonically with time. In the above basic LVQ1 it is recommended that α should initially be smaller than 0.1 Kohonen [2]; linear decrease in time is used in this research when we used optimized LVQ (OLVQ).

II. The optimized-learning-rate LVQ1 (OLVQ1)

The basic LVQ1 algorithm is now modified in such a way that an individual learning rate $\alpha_i(t)$ is assigned to each m_i . and that can approximately be optimized for quick convergence. Then the following discrete-time learning process is obtained. If c is defined by equation. (1). Then

$$m_c(t+1) = m_c(t) + \alpha_c(t)[x(t) - m_c(t)] \quad \text{if } x \text{ is classified correctly}, \quad (5)$$

$$m_c(t+1) = m_c(t) - \alpha_c(t)[x(t) - m_c(t)], \quad \text{if } x \text{ is classified incorrectly}, \quad (6)$$

$$m_i(t+1) = m_i(t) \quad \text{for } I \neq c. \quad (7)$$

Next, the problem of whether the $\alpha_i(t)$ can be determined optimally for fastest possible convergence of (3) was addressed. If equation (6) is expressed in form of;

$$m_c(t+1) = [1 - s(t)\alpha_c(t)]m_c(t) + s(t)\alpha_c(t)x(t) \quad (8)$$

where:

$s(t) = +1$ if the classification is correct and

$s(t) = -1$ if the classification is incorrect,

it is noted here that $m_c(t)$ is statistically independent of $x(t)$. It might also be noted how optimal the statistical accuracy of the learned codebook vector values, if the effects of the corrections made at different times are of equal weight. Notice that $m_c(t+1)$ contains a "trace" from $x(t)$ through the last term in equation (8), and "traces" from the earlier $x(t')$, where $t' = 1, 2, \dots, t-1$ through $m_c(t)$. The (absolute) magnitude of the last "trace" from $x(t)$ is scaled down by the factor $\alpha_c(t)$, and, for instance, the "trace" from $x(t-1)$ is scaled down by: $[1-s(t)\alpha_c(t)]\alpha_c(t-1)$.

It was made certain that these two scaling factors are identical:

$$\alpha_c(t) = [1 - s(t)\alpha_c(t)]\alpha_c(t-1) \quad (9)$$

Now if the condition is fulfilled to hold for all t , it can be shown that the "traces" accumulated up to time t from all the earlier x will be scaled down by an equal amount at the end, and thus the "optimal" values of $\alpha_i(t)$ are determined by the recursion.

$$\alpha_c(t) = \frac{\alpha_c(t-1)}{1 + s(t)\alpha_c(t-1)} \quad (10)$$

After we implemented the optimal version of LVQ in our system, it became obvious how equation (10) really provides fast convergence. However, we noticed that precaution must be made when choosing $\alpha_c(t)$ so that it does not rise above the value 1, the interesting about learning program OLVQ1 that it never allows any α_i to rise above its initial value. It is proved that the initial values of the α_i can be selected rather high, say, 0.3, whereby learning is considerably speeded up, especially in the beginning, and the m_i quickly finds their approximate asymptotic values.

3 Algorithm Parameters and Implementation

After the brief theoretical background of OLVQ algorithm discussed in previous section we need to initialize the network and implement the algorithm utilizing all parameters come with algorithm package provided by Kohonen [1,2]. The OLVQ algorithm has been implemented in our Arabic OCR system. Classification is performed in two major steps; training and classifying.

In training stage number of parameters like; input and output files, learning rate $\alpha_{i(t)}$, number of codebook vectors, code-vectors for each class (i.e. basic-shape character), number of iterations, number of k-NN and learning rate function, need to be tuned for LVQ network so that optimal recognition rate is obtained.

4 Classifier Assessment

The codebook file contains codebook vectors that approximate to the samples of the input vectors, one codebook vector being assigned to each sample. A set of fourteen Arabic character basic-shapes is used to test the classifier, the characters used are; ي (ya), و (waw), س (ssad), ش (seen), ر (ra), م (meem), ل (lam), ك (kaf), ح (hha), ه (ha), د (fa), ب (dal), ئ (ba) and ئ (ain) respectively). Number of samples for each character is 21 samples. Two different data sets of these samples are prepared, one for training and the other for testing. The set of features of each character is referred to as a *vector*. Each input vector has a dimensionality of 12 floating point numbers followed by the class (character) label. A number of 21 samples of each character (total of 294 input vectors) are stored in a training data file called *Features1.dat*, and the same numbers are stored in a testing data file called *Features2.dat*. A codebook is initialized with a total number of 200 code-vectors. The number of entries selected for each class (character) is check and the medians of the shortest distances are calculated.

It has been noticed that the recognition accuracy depends on the number of codebook entries allocated to each class, and the best distribution of the codebook vectors is not easily predicted. In this experiment we used the method of iteratively balancing the medians of the shortest distances in all classes. Balancing of the medians is achieved by calculating the medians of the shortest distances for each class first, then correcting the distribution so some of those classes which have distances greater than the average entries are added, and some of those classes in which the distance is smaller than the average are deleted.

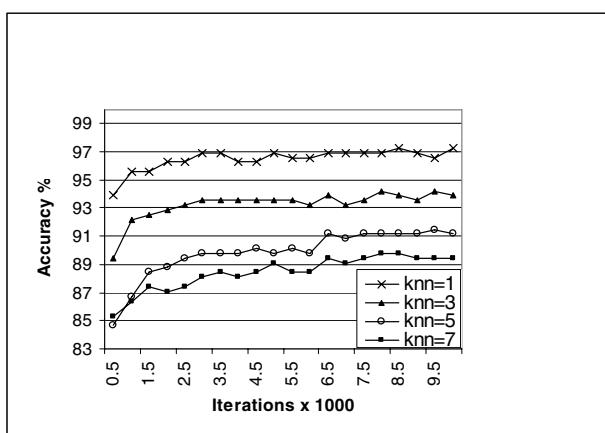


Fig. 3. The Effect of Iterations and knn on Recognition Accuracy

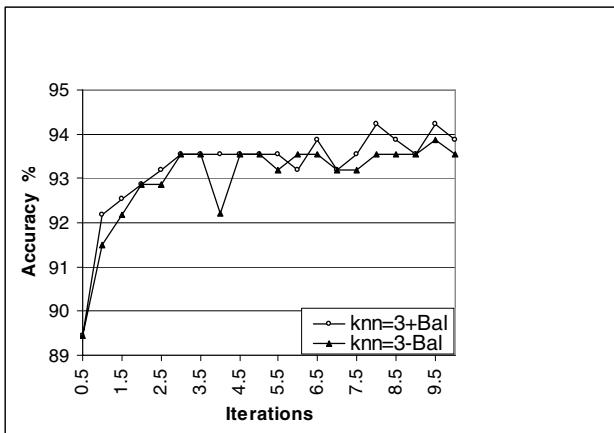


Fig. 4. The Effect of medians balancing on overall accuracy, \pm means with or without balancing

Now the codebook has been initialized and ready for learning. The codebook is trained by optimized-LVQ, which has been described by Kohonen [3] as the fastest and most robust of all the Learning Vector Quantization algorithms. Number of training run length (training iterations) are tried, and the best found to be 7000 times which can achieve optimal accuracy as shown in Figure 3. The Effect of medians balancing on overall accuracy is shown in Figure 4.

Classification has been performed in two different strategies, in first strategy, we use one classifier for all 53 Arabic CBSs in training and testing phases, in second strategy we use three classifiers and three subsets of 53 Arabic CBSs. The three subsets of Arabic CBSs are; Ascenders (characters above base line), Descenders (characters below base line) and Embedded (characters inside base area), and the classifier is trained for each of them individually.

4.1 First Strategy

In this strategy, as we mentioned earlier, all 53 Arabic CBSs are used as one sets for training and testing phases for one LVQ network. A total of 21 samples of each CBS are used in learning and testing steps, 11 samples of each CBS are taken as training set whereas the rest 10 samples of each CBS are used as testing set. Each CBS of training and testing sets are translated into code-vector by features extraction module and stored into two different data file, one data file contains code-vectors of all CBSs belong to training set, the other data file contains code-vectors of all CBSs belong to testing set. We started by initialization of codebook-vectors, then the medians of the shortest distances between the initial codebook vectors of each class are computed (balancing), then we trained the LVQ network by specifying different combinations of learning parameters like; input and output files, learning rate $\alpha_i(t)$, number of codebook vectors, code-vectors for each class (CBS), number of iterations and number of k-NN. Finally we tested the network and monitor the recognition accuracies for

each individual class (CBS) as well as the classifier overall recognition rate. The overall recognition accuracy of this classifier is 83.2 %.

4.2 Second Strategy

This strategy is based on numbers of hypothesis and notions we started with. First, if we take all basic shapes (53 CBSs) the possibility of finding two or more of these basic shapes being morphologically close to each other is not ignorable. Second, the more the number of classes to be classified the more the complication of network design and programming are claimed. Third, tasks subdivision facilitates parallel programming, fast debugging and system development. Based on what we mentioned here we decided to use three classifiers instead of one, we divided the 53 CBSs into three groups ascenders, descenders and embedded, and accordingly the three classifiers are; ascenders classifier, descender classifier and embedded classifier.

The same training and testing procedures followed in case of first strategy is applied for each individual classifier in this strategy. Each individual classifier is trained and tested with its correspondent dataset. The overall recognition accuracies of ascenders, descenders and embedded classifier are 92.21 %, 88.24 % and 83.2 % respectively. The average recognition rate of the three classifiers is 89.1 % which is higher than the recognition rate of the first strategy classifier (which is 83.83 %). It was concluded that as the number of CBS classes' increases the recognition accuracy decreases, and this reassure our presumption of dividing the CBSs set into three categories and using three classifiers, one classifier for each category.

It has been noticed that there are two fundamental sources of misclassification for our classifier. The first one is the low number of available CBSs samples for training and testing. The second is the intrinsic ambiguity in Arabic cursive characters. Characters in a handwritten word often have flourish and ligature strokes which generally do not appear in independently written characters. These extra strokes are hard to be generalized in common prototypes, since they are usually unique to the writing style. The training character set can not cover all the variations of these strokes even if pre-isolated characters from word images are used for training because of the uniqueness of these extra strokes.

5 Conclusion

We have presented a new method for Arabic handwritten character recognition using a reliable features extraction module and LVQ neural network technique as classifier. LVQ is one of the best clustering techniques. It has been shown that the proposed method is more effective than the conventional matching methods used in OCR systems. This method is robust with regard to geometrical variation, but very sensitive to topological variations such as the presence of spurious small branches, loops opening or strokes closing etc. The average recognition rate using three classifiers (one for each subset of Arabic Character Basic Shapes CBSs) is higher than the recognition rate using one classifier for all 53 Arabic Character Basic Shapes CBSs.

References

1. Cao, J., Ahmadi, M., Shirdhar, M.: Recognition of handwritten numerals with multiple feature and multistage classifier. *Pattern Recognition* 28(2), 153–160 (1995)
2. Kohonen, T.: New developments and applications of self-organizing maps Neural Networks for Identification. In: Proceedings of International Workshop on Control, Robotics, and Signal/Image Processing, pp. 164–172 (1996)
3. Kohonen, T., Barna, G., Chrisley, R.: Statistical pattern recognition with neural networks: benchmarking studies. In: IEEE International Conference on Neural Networks, vol. 1, pp. 61–68 (1988)
4. Yu, D., Yan, H.: Separation of single-touching handwritten numeral strings based on structural features. *Pattern Recognition* 31(12), 1835–1847 (1998)
5. Liwei, W., Xiao, W., Jufu, F.: On image matrix based feature extraction algorithms. *IEEE Transactions on Systems, Man and Cybernetics* 36(1), 194–197 (2006)
6. Kenneth, E.H., Deniz, E., Kari, T., Jose, C.P.: Feature Extraction Using Information-Theoretic Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(9), 1385–1392 (2006)
7. Menasri, F., Vincent, N., Cheriet, M., Augustin, E.: Shape-Based Alphabet for Off-line Arabic Handwriting Recognition. In: Ninth International Conference on Document Analysis and Recognition, ICDAR, vol. 2, pp. 969–973 (2007)

A Complete Pyramidal Geometrical Scheme for Text Based Image Description and Retrieval

Guillaume Joutel, Véronique Eglin, and Hubert Emptoz

LIRIS UMR CNRS 5205 – INSA Lyon, 69621 VILLEURBANNE Cedex
{guillaume.joutel; veronique.eglin@insa-lyon.fr}

Abstract. This paper presents a general architecture for ancient handwriting documents content description and retrieval. It is based on the Curvelets decomposition of images for indexing linear singularities of handwritten shapes. As it belongs to the Wavelets family, its representation is used at several scales of details. The proposed scheme for handwritten shape characterization targets to detect oriented and curved fragments at different scales: it is used in a first step to extract visual textual interest regions and secondly to compose a cross-scale signature for each handwritten analyzed samples. The images description is studied through different kinds of deformations that show the efficiency of the proposition for even degraded and variable handwriting text. The complete implementation scheme is validated with a content based images retrieval (CBIR) application on the medieval database from the IRHT¹ and on the European 18th century correspondences corpus from the CERPHI.

1 Introduction

1.1 Digital Paleography and Image Analysis

In this work, we are interested in digitized Middle-Age (composed by copyists' texts from the 9th to the 15th century) and Humanistic manuscripts (essentially composed by authors' drafts from the 18th and 19th century) analysis. It is dedicated to palaeographers and historians to help them in their everyday work of manuscripts dating, expertise and authentication, see figure 1. One of the primary difficulties faced by palaeographers is the classification and identification of hands, and this is an area which has already received a good deal of attention in other disciplines. Specifically, the community of forensic document analysts have been working for several years now to develop computer-based systems for retrieval, identifying and classifying modern handwriting, and this begs the question whether such work can be applied to medieval writing as well. In that context, we need an original approach to answer the opposite questions raised by those two kinds of documents from the *medieval* and the *humanistic* collections. In the first case, the objective is to find writer independent and

¹ This study has been supported by two projects: the ANR GRAPHEM 2007-2010 project between the IRHT (Institut de Recherche en Histoire des Textes), the Ecole des Chartes from Paris and the LIRIS and the regional Cluster 13 Project: “Inheritage, culture, Creation”, 2005-2009 in collaboration with the CERPHI of Lyon (Centre d’Etude en Rhétorique, Philosophique et Histoire des Idées).

style dependent primitives for Middle-Ages manuscripts page images retrieval, while for humanistic manuscripts, the objective is to find writer dependent primitives for an identification task. The answers to those objectives are complex. We propose in this paper a contribution to this challenge that must be considered as a tool kit for historians and medievalist palaeographers to help them in their work of manuscripts dating, expertise and authentication. Due to the fact that palaeographers judge that the *curvature* and the *orientation* are two fundamental dimensions of handwritings, we have searched a way to compute them on the shapes for both collections. To compute those dimensions, we have developed a methodology that is sensitive to variations at the frontiers of shapes, that pays attention to the evolution of these criteria at different scales, that reveals the variability of shapes and their anisotropy, that is robust to disturbed environments (presence of disturbed backgrounds, of partial shapes...) and finally, that doesn't require prohibitive costs of treatment nor great storage volume for each analyzed page image. Our choice relates to a redundant multi-scale transform: the Curvelets transform that is more robust than wavelets for the representation of shapes anisotropy, of *lines segments* and *curves* in the images. The proposed solution lies for a great part on an original indexing and navigation system that contributes to different objectives: handwritten documents retrieval from the same writer, writers' identification, middle-age handwriting clustering.

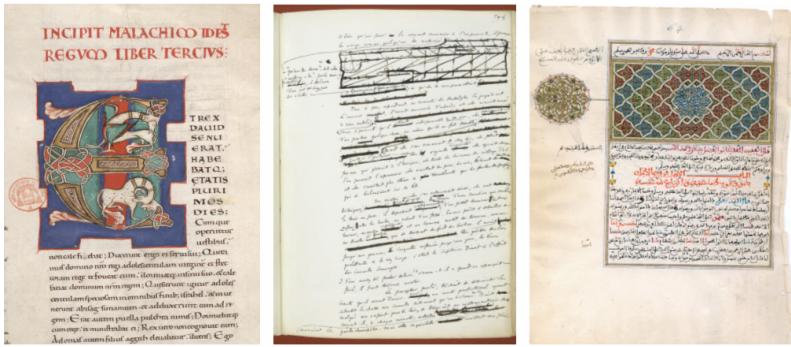


Fig. 1. Complex handwriting documents images from different periods (13th and 18th century)

1.2 A Methodology Based on Coarse Interest Regions Selection and Fine Shape Description

Our CBIR system is based on geometrical multiscale images decomposition (the Curvelets) embedded in the different stages of the architecture. The figure 2 presents the system architecture into two main parts.

The first part presents an off-line Curvelet based decomposition that lies on the exploitation of the distributions of Curvelet strong coefficients through different scales: the lowest scales for the estimation of interest text areas, the middle and highest scales for the computation of digital signatures of homogeneous text fragments. The digital signatures are the 2D representations of all couples of values (curvature, orientation) obtained by the Curvelets decomposition. Three applications have then been investigated by exploiting this off-line decomposition:

- The *writing styles clustering* that is based on the production of a cross-scale features vector for each handwritten samples and that contains a set of statistical measures from the different multiscale coefficients distributions. This clustering is based on an unsupervised classification scheme that groups palaeographical handwritten styles into visual coherent handwriting families.
- The *spotting* and the *shape retrieval by similarity* that lies on the analysis of the curvature and the orientation from the finest scale of the decomposition of the image.
- The *CBIR* that is also based on the finest scale analysis and on the production of an unique and characteristic digital signature for each handwritten samples.

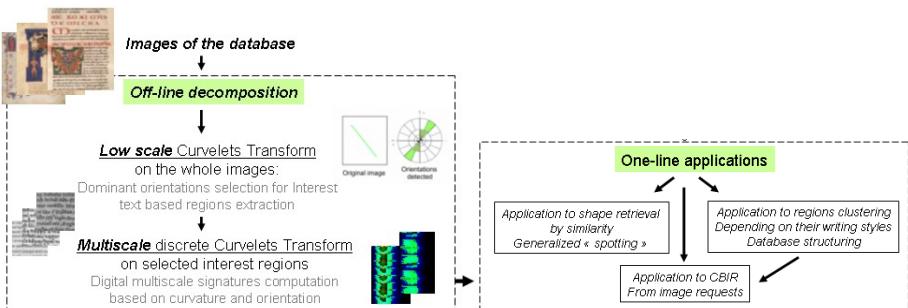


Fig. 2. Pyramidal architecture for the handwriting characterization and on-line applications

In this paper, we focus on the application of CBIR that is based on a direct comparison between a request image, considered by its digital signature and all off-line signatures for the rest of the database images. The similarity criterion is a weighted value of correlation between the request signature and each individual signature of the images of the database.

1.3 Content Based Information Retrieval for Handwriting Images Databases

The simplest scenario of content based image retrieval is the one of global example-based search: the user chooses an image example and the system determines the images of the base with the most similar visual appearance. The principle of this approach has been established by Ballard in 1991, [Ballard]. It is the fundamental principle of a lot of systems that deal with natural images, like QBic from IBM (Flickner 1995), PhotoBook from the MIT (Pentland 1994), MARS (Multimedia Analysis and Retrieval System), PicToSeek, Ikona and the KIWI system, [10]. All those systems are sharing the same kinds of features for a global content based document images retrieval (colors, shapes and textures). In the specific field of CBIR dedicated to textual (printed or handwritten) document images, different specific (structural, geometrical or statistical) features and similarity measures have also been proposed, [9], [11]. But all of them have a common point: they are focused on the

same property of images: the information is concentrated in linear shapes and their contours.

2 Different Ways to Characterize Handwritten Shapes

2.1 Related Works

The need for searching scanned handwritten documents are involved in application such as collections dating, works genesis, erudite study, critical edition, documents authentication... Recently Srihari et al. in [15-16] have realized the importance of handwritten document retrieval and have presented their retrieval system that is dedicated to forensics applications such as writer identification. Different distances are currently used to access the best matching between different set of handwriting samples, [7]. Discriminability of a pair of writing samples based on similarity value can be observed by studying their distributions when the pair arises from either the same writer or from different writers. Generally, in most writers' classification approaches, authors try to produce a set of exhaustive graphical characteristics which can efficiently describe all handwritings families. In handwriting classification which is our main goal, the most closely related works are [8] and [16]. In [8], authors propose an analysis of the variability of handwritings on the bases of two kinds of observations: firstly the thickness of the tracing and the spatial density of characters and secondly the successive directions of the tracing. This work has been led on contemporary documents only. The only work relative to the analysis of the Middle-ages documents referenced in [1] proposes classifications procedures which are today debatable by palaeographers. Some recent studies have tried to develop more robust approach for a non supervised classification. In [11] for example, authors are interested in ancient Latin and Arabic manuscripts of the Middle-Ages before the emerging of the printing. They made the choice to analyze statistically the whole image of a manuscript and measure globally all patterns. This approach should guarantee the independency from the text content, the writer's personal style; the language used and the letters frequencies. It is difficult to pretend to be exhaustive in the description of handwritten shapes for the retrieval, so it is essential to work with expert users who are able to validate the measures that appear to be the most relevant. The signification degree that is assigned by the user could also guide the system to create a suitable distance between writing classes. In any cases, each computed measure should be evaluated in relation with all others. In this case, a Principal Component Analysis (PCA) considerably reduces the dimensionality of micro-features vectors, [13]. Within this kind of generic approach, it is possible to classify handwriting samples into visual distinct style classes. General methods are either based on the consideration of local particular graphemes [3] or on a too macroscopic and general characterization that is not always efficient for a complete writers' authentication, [5], [11]. In our proposition, we lead a Curvelets based analysis for a designed decomposition in curvature and orientation which have a real sense for palaeographers. Our contribution differs from the cited works because it is dedicated to both Middle-Ages and contemporary handwritings

documents. Moreover it uses a geometrical multiscale decomposition: the Curvelets, which has never been used in this context.

2.2 A New Interesting Way: The Geometrical Decompositions

In the handwriting images, the discontinuities points are essentially gathered on the shapes' outlines that contain the maximum of the visible information. It is an information that can be analyzed both globally (orientations of outlines, presences of particularly striking and redundant curvatures), and locally (the discontinuities are situated in very sensitive locations in shapes contours where experts pay generally their attention to recognize or identify a writing).

A mixed approach that combines global and local shapes description seemed to be an interesting way for handwriting images analysis. But we have noticed that all decompositions can not decorrelate both local and global images properties with the ability to localize all discontinuities in the image space. Fourier Transform is a powerful tool for analysing the global regularity of a function but it does not allow recognizing discontinuities points. It doesn't allow a local shape analysis and doesn't take advantage of multiscale correlations that exist in lines based images. Orthogonal wavelets provide frames that allow sparse representation (with only few coefficients) of images but present also a lack of directional characteristics for regular curves description and geometry. We are precisely interested here in the families of redundant transforms which allow to decompose the image into directed sub-bands and to capture the geometry in the shapes contours by the agglomerates of coefficients. We can give as example, the X-lets (*i.e.* contourlets, bandelets, complex wavelets and oriented wavalets). Among this family, we have chosen one specific X-lets: the Curvelets that have been introduced by Candès. They offer a set of the properties of directionality and good joint location of the shapes geometry. Ridgelets and Curvelets transforms that we defend here have been imagined to bypass wavelets disadvantages.

2.3 The Curvelet Transform - From Continuous to Discrete

General Properties of the continuous Curvelet Transform. The Curvelet transform is a multiscale multi-orientation transform with indexed atoms in location, scale and direction. The figure 3 shows the anisotropic Curvelet transform of Candès [6] of an image of our palaeographical database. On this figure, the transformation contains 4 scales with all coefficients, each one being of the same size as the original image. The combination of all images with all missing scales (not represented here) reproduces with redundancy the original image. A study of the Curvelets coefficients properties, shows that the high valued coefficients tend to gather around the objects corners and contours in the image, what seems to be natural because there is a strong dependence between Curvelets coefficients from a scale to the other and for a given neighbourhood for various given scales and orientations. These dependences have been quantitatively verified by measuring the mutual information, [4].

Conceptually, the Curvelet transform is a multi scale nonstandard pyramid because Curvelets have geometric features that set them apart from wavelets and the likes.

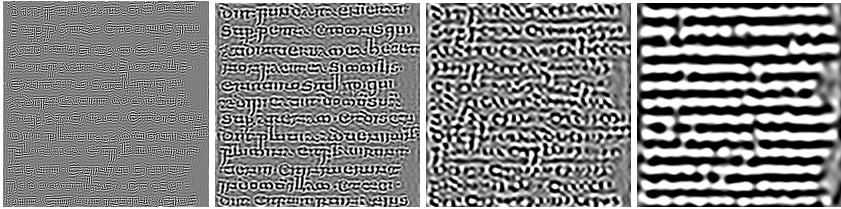


Fig. 3. Curvelet Transform of an image of the Palaeogeographical database (IRHT-Paris)

Curvelets obey a parabolic scaling relation which says that at scale 2^{-j} , each element has an envelope which is aligned along a ridge of length $2^{-j/2}$ and width 2^{-j} . Mathematically, if one works in \mathbb{R}^2 , we have first to consider a radial window $W(r)$ and then an angular window $V(t)$ where r and t are polar coordinates in the frequency domain. These are both smooth, nonnegative and real-valued, with W taking positive real arguments and supported on $r \in [1/2, 2]$ and V taking real arguments and supported on $t \in [-1, 1]$. These windows will always obey the admissibility conditions:

$$\sum_{j=-\infty}^{\infty} W^2(2^{-j}r) = 1, \quad r \in (3/4, 3/2); \quad \sum_{\ell=-\infty}^{\infty} V^2(t - \ell) = 1, \quad t \in (-1/2, 1/2)$$

Then, for each $j \geq j_0$, a frequency window U_j is defined in the Fourier domain by

$$U_j(r, \theta) = 2^{-3j/4} W(2^{-j}r) V\left(\frac{2^{\lfloor j/2 \rfloor} \theta}{2\pi}\right).$$

where $\lfloor j/2 \rfloor$ is the integer part of $j/2$. The support of U_j is a polar wedge with the support of W and V , the radial and angular windows applied with scale-dependent window widths in each direction.

Procedural definition and implementation of the discrete version. The proposed Curvelets decomposition uses a combination of reversible operations: a sub-bands decomposition of the original 2D signal followed by a regular and normalized partitioning of images for each sub-band and then the application of a local Ridgelets transform on each partition.

We now describe a strategy for realizing a digital implementation of the Curvelet transform. The proposed implementation of the Curvelets decomposition has been suggested by Candès and Starck and constitutes the most common version. The successive simplified steps of the algorithm are listed below:

- 1/ Do the FFT of the initial image I
- 2/ Resample the frequential Cartesian space into recto-polar coordinates
- 3/ For each radial indexed n line do the inversed Fourier Transform: so as to obtain the Radon Transform $\{\text{Rad}(n, t_n)\}_{n=1..N}$ for each n angle and then realize the wavelet 1D transform in the spatial domain.
- 4/ The inverse Ridgelet transform realizes all successive inversed transforms on each step.
- 5/ Initialize the subband decomposition for the Curvelet transform step.
- 6/ Compute a local Ridgelet transform on each subband.

In practice, we have privileged the new wrapping-based curvelet transform proposed by (Candès, 2002) that simplifies the calculation and increases the redundancy of the coefficients

3 Off-Line Decomposition for the CBIR Application

Interest textual regions selection. In textual regions of document images, lines orientations are globally regular near to the horizontal. This orientation can be detected at a low scale : the map of textual information is revealed by high Curvelets coefficients while other coefficients have been ignored. Of course, at this resolution of image, we have lost a part of precision in this detection that has been easily compensated by the detection of the *background map* that has been introduced in Ramel's work in [12].

Signatures construction, evolution and comparison. The Curvelet transform gives us an analysis of pixels for several scale and several orientations. We only focus here on the highest scale of the decomposition and only cumulate information relative to shapes contours. The principle of curvature estimation is described as follows: one pixel on a curve can be potentially detected in several orientations depending of the curve. Each Curvelets coefficient corresponding to this pixel in a detected orientation is then compared to all other detected orientations of the same pixel. We only retrieve the more significant coefficients values. Finally, the number of significant orientations gives us an evaluation of the overall curvature of a pixel, as it has been presented in [2] by J-P Antoine. From the orientations recovery and curvatures computation, we define a signature for each handwriting fragment. The signature is defined as the matrix of couple occurrences (curvature, orientation). Because we do not want to normalize our images before the decomposition, we have searched a way to normalize the signature. To do so, we keep in the normalized signature the ratio between every value in the original signature and the global amount of information of this signature. The only difficulty with this definition is that some coordinates in the signatures are common to almost all handwritings (presence of high values in the horizontal and vertical directions): we compensate this phenomenon by reversing ratios in the signature, i.e. by accentuating weakly represented couples (orientation, curvature). Figure 5 presents an example of matrix plotted as a pixel image for a Middle-Ages and a 18th author handwriting image. In our retrieval system, the distance between the queried signature and each signatures of the database is calculated using a normalized correlation similarity. The similarity measure $S(X, Y)$ between two image X and Y is defined as follows: $S(X, Y) = \text{Cov}(X, Y) / \sigma_x \sigma_y$, where $\text{cov}(X, Y)$ is the covariance between X and Y and σ_x and σ_y the standard deviations. This simple measure is rigid enough to really match datas with a link. Some distances such as the inter-correlation or the dynamic time warping have been tested but better results are obtained with the correlation.

In order to know how our signature reacts to several possible deformations in a handwriting manuscript, we have led five kinds of tests: the horizontal and vertical stretching of text, the resolution changing, the letters' dimensions zooming and the lines' spaces changing. To have significant results, we have chosen to keep the same image dimensions after each deformation.

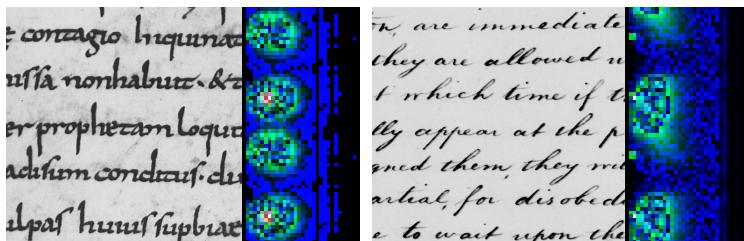


Fig. 5. Signature of a Middle-Ages handwriting samples of the IRHT database and on a fragment extracted from the Georges Washington’s digital collection

The only general conclusions are presented here. Firstly, studying (vertical or horizontal) stretching, the signature's behaviour is exactly what we were expecting: i.e. a general shifting of values of couples (curvature, orientation) towards the orientation of the stretching (0° for horizontal and 90° for vertical), a general decreasing of curvatures and a great fall of the correlation coefficients between the original image and its successive deformations. In fact, the horizontal and the vertical directions are the two main dimensions in Latin handwritings. So modifying those dimensions led to great changes in the signatures design. Concerning the resolution and the letters' dimensions changes, we can take advantage on the presence of different scales in the decomposition of the Curvelets that stabilize the signature. The Curvelet transform offers the possibility to choose the best scale for the signature design depending on the initial image resolution and characters size. Lastly, modifications of the spaces between lines entail real consequences on the signature only when the amount of information is no more comparable. In an image with about ten lines of medieval text, if one removes or adds two or three lines, the rate of recognition is maintained.

Another test was to sign the background of a document without any text and the entire document separately. Correlation between the two signatures was clearly no significant (about 0.1)

4 The CBIR Application: Experimentation and Quantitative Evaluation

In this part, the test setup and the experimental results obtained for the image retrieval task are described. The images have been divided into two groups: one group is composed by known handwriting images and the second group is composed by queried images for testing. We have tried to test our retrieval system on two databases (over 800 images): the humanistic and the Middle-Ages databases (composed by manuscripts of the database of the IRHT and of European 18th century correspondences corpus (CERPHI)). In retrieval system, the performance is subjective and its evaluation is difficult, especially for the Middle-Ages base where we do not have precise information concerning Middle-Ages European handwriting dating process. The set of known images is selected and the image retrieval process is carried out against this set of known handwritings. In each case, we compute the average precision and the recall values for each database.

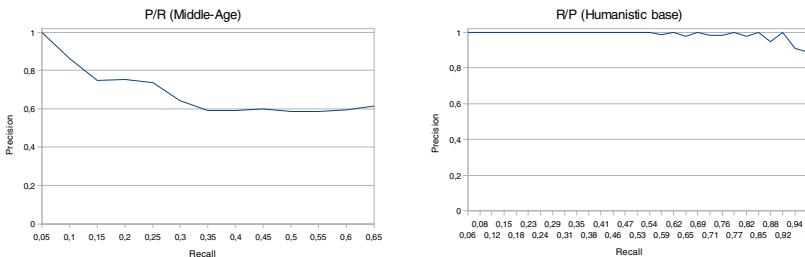


Fig. 6. Precision/Recall curves for the Middle-Age (on the left: 6.1) and the Humanistic (on the right: 6.2) database

As we can see on figure 6.1 the P/R curve tends to stabilize around a precision rate of 0,6 which is not a very high value but we have proceeded by unfavourable deductions that consists in counting as false answers all images that are not surely recognized as being similar to the request. This is due to the lack of certain ground truth about classification of the middle age database. The classification of our base is still in progress by palaeographers of the IRHT. On the other hand, on the figure 6.2, we can see that our system has exactly the same behaviour that we expected. In fact those rates are maybe the best that we can have, compared to other CBIR systems.

To have a better evaluation of our system, we have used a F-measure which combines the precision and the recall values ($F=2P.R/(P+R)$) and provides a single measure of the retrieval accuracy: it is the Harmony mean of precision and recall value. For the humanistic base, in the top 15 ranks, the F-measure is 96,61 (P=1 and R=93). This implies that considering the top 15 results supports high precision and recall measure. For the Middle-Ages base this value is 51,43.

5 Conclusion

We can say with confidence that Curvelets Transforms can be used as a general technique for feature detection. In the specific case of handwritings that we study, the hypothesis that curvature and orientation are fundamental dimensions of shapes seem to be validated by our CBIR system and its recall values. The system is currently integrated to the platform of the Graphem Project dedicated to palaeographer experts. The two other on-line applications that are also based on the Curvelet geometric decomposition (spotting and writing styles clustering) are currently under development.

References

1. Aiolfi, F., Simi, M., Sona, D., Sperduti, A., Starita, A., Zaccagnini, G.: SPI: a System for Palaeographic Inspections. *AIIA Notizie* 4, 34–38 (1999)
2. Antoine, P., Jacques, L.: Measuring a curvature radius with directional wavelets. *Inst. Phys. Conf Series*, pp. 899–904 (2003)
3. Bensefia, Heutte, L., Paquet, T., Nosary, A.: Identification du scripteur par représentation graphémiques. In: *CIFED 2002*, pp. 285–294 (2002)

4. Boubchir, L., Fadili, J.: Bayesian Denoising Based on the MAP Estimator in Wavelet-domain Using Bessel K Form Prior. In: Proc. of IEEE ICIP 2005; the IEEE International Conference on Image Processing, Genoa, Italy, September 11-14, vol. I, pp. 113–116 (2005)
5. Bulacu, M., Schomaker, L.: Writer Style from Oriented Edge Fragments. In: Petkov, N., Westenberg, M.A. (eds.) CAIP 2003. LNCS, vol. 2756, pp. 460–469. Springer, Heidelberg (2003)
6. Candés, E., Donoho, D.: Curvelets: A Surprisingly Effective Nonadaptive Representation of Objects with Edges. In: Schumaker, L. (ed.) Curves and surfaces filtering, Vanderbilt University Press (1999)
7. Cha, S.H., Srihari, S.: Multiple Feature Integration for Writer Verification. In: IWFHR VII, pp. 333–342 (2000)
8. Crettez, J.-P.: A set of handwriting families: style recognition. In: ICDAR 1995, pp. 489–494 (1995)
9. Journet, N., Mullot, R., Ramel, J.Y., Eglin, V.: Ancient Printed Documents indexation: a new approach. In: ICAPR 2005. Third International Conference on Advances in Pattern Recognition, Pattern Recognition and Data Mining, Bath, United Kingdom, pp. 513–522 (2005)
10. Loupias, E., Bres, S.: Key Points based Indexing for Pre-attentive Similarities: The KIWI System. Pattern Analysis and Applications, Special Issue Image Indexing (summer 2001)
11. Moalla, I., LeBourgeois, F., Emptoz, H., Alimi, A.M.: Contribution to the Discrimination of the Medieval Manuscript Texts. In: Bunke, H., Spitz, A.L. (eds.) DAS 2006. LNCS, vol. 3872, pp. 25–37. Springer, Heidelberg (2006)
12. JY. Ramel
13. Shen, C., Ruan, X.G., Mao, T.L.: Writer identification using Gabor, vol. 3, 2061–2064 (2002) [8, 13, 24]
14. Swain, M.J., Ballard, D.H.: Color Indexing. International Journal of Computer Vision 7(1), 11–32 (1991)
15. Zhang, B., Srihari, S.N.: Binary vector dissimilarity measures for handwriting identification. In: Document recognition and Retrieval, SPIE, vol. 5010, pp.28–38 (2003)
16. Zhang, B., Srihari, S.N.: Word image retrieval using binary features. In: Document recognition and Retrieval, SPIE, vol. 5296, pp.45–53 (2003)

Local Orientation Extraction for Wordspotting in Syriac Manuscripts

P. Bilane, S. Bres, and H. Emptoz

LIRIS, INSA-Lyon, F-69621, France
{petra.bilane, stephane.bres, hubert.emptoz}@insa-lyon.fr

Abstract. This paper presents a contribution to Word Spotting applied for digitized Syriac manuscripts. The Syriac language was wrongfully accused of being a dead language and has been set aside by the domain of handwriting recognition. Yet it is a very fascinating handwriting that combines the word structure and calligraphy of the Arabic handwriting with the particularity of being intentionally written tilted by an angle of approximately 45°. For the spotting process, we developed a method that should find all occurrences of a certain query word image, based on a selective sliding window technique, from which we extract directional features and afterwards perform a matching using Euclidean distance correspondence between features. The proposed method does not require any prior information, and does not depend of a word to character segmentation algorithm which would be extremely complex to realize due to the tilted nature of the handwriting.

Keywords: Word Spotting, orientation features, directional roses.

1 Introduction

The Syriac language belongs to the Aramaic branch of the Semitic languages. The oldest Syriac manuscripts can date back to the 1st century AD, however the literature itself flourished from the 3rd century AD onward, then began to decline in the 7th century in the face of Arabic culture [2].

The documents we are interested in are old manuscripts. Most of the time these documents present much degradation that can be interpreted as noise in the context of text extraction and recognition. If we take into account the variability of the handwriting and the fact that the segmentation of the text into letters is most often not possible, we understand why classical OCR is useless. The word is then the smallest element we can identify.

2 Related Work

Very few are the people who launched themselves in the study of Syriac manuscripts. Besides the works of William Clocksin [1], no previous work has been published on Syriac handwriting recognition. Different approaches exist for handwriting recognition

in historical manuscripts. In our previous work, we were interested in global information for document classification based on handwriting style [4]. In this paper, we focus on word spotting.

Old documents can be treated as they are or a pre-processing can be performed, like binarization to highlight the text [5][6]. Most of the approaches focus on words and not letters [7][8][10][14], because of the problems of segmentation on manuscripts data. For most of the authors, a word level approach is better than a letter level approach. In this domain of old documents study, there are a lot of existing documents. It is not always that easy to have access to them and thematic studies on documents coming from a precise origin are especially interesting because they can take advantages of some constant characteristics of the database. In this paper we focus on Syriac documents. Other researches were made on the famous French writer Flaubert corpus [9] for layout extraction using Markov fields, Manmatha [8][10] targeted manuscripts from the Georges Washington's collection and Leydier targeted Medieval Latin manuscripts [7]. The purpose of their works is to extract and recognize words using description based on computed features. Terasawa et al. [11] also performed word spotting inspired by an Eigen space method [11] or gradient [12]. In this case, word signatures are extracted from sliding windows. The relative levels of the gradient in the 8 main directions are computed in these sliding windows. This leads to features that are robust to scale changes. To overcome the morphological differences between the words, the matching is performed using a Dynamic Time Warping (DTW) algorithm. DTW is also used in [1] to match whole words. Another segmentation-free approach which uses HMMs and statistical language models for handwritten text recognition is described in [13].

3 Proposed Method

The method that we propose consists of an eliminatory process. First we start with a preprocessing phase to select sliding windows of interest, unlike Terasawa et al. [10] who took into consideration all extractable windows. The elimination begins at this step: windows that do not respond to certain criteria are eliminated (the elimination process is further detailed in the following sub sections). Regions of interest are then detected. Afterwards, saliency coefficients of directional roses are extracted from sub windows within the regions of interest, and are matched to those extracted from the query word image using a point to point correspondence of the Euclidean distance.

3.1 Preprocessing

The material that we will be working on consists of Karshuni manuscripts that are written with the Serto calligraphy. They were supplied to us by the Central Library of the Holy Spirit University of Kaslik in Lebanon. These manuscripts date from the beginning of the 19th century and were digitized with a resolution of 300 dpi. Figure 1 shows a sample from these manuscripts.

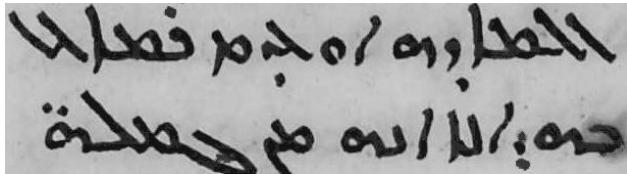


Fig. 1. Sample image from a Karshuni manuscript written with the Serto calligraphy

The manuscripts that we have contain mainly three color categories, the background which is a light yellow tone, the main text which is dark brown or rather black, and the emphasized words which are written in red. The preprocessing phase consisted converting the document in grey level preserving the contrast between the text and the background. A skew detection and correction of the line is then performed using the Hough Transform. An upper and a lower baseline detection using horizontal projections are finally performed (see Figure 2).

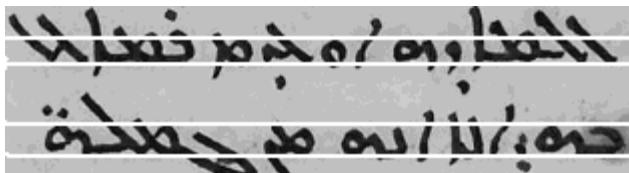


Fig. 2. Example of an upper and a lower baseline detection

3.2 Selective Sliding Windows

Once the upper and lower baselines are detected, the text line is divided into three regions (upper, middle, and lower). In each of these regions a sliding window of size 32x32 pixels is passed along from left to right at a step of one pixel. Afterwards, an analysis of the content of the window is conducted: first we reject the windows having less than a certain minimum density of black pixels; then we study the movement of the center of gravity of the black pixels. We only keep the window having a center of gravity that moved significantly along the x axis compared to its predecessor; in a third elimination, windows that cover more than half their predecessor are eliminated.

The size of the sliding windows and the allowable covering rate were chosen while taking into consideration the image resolution (300 dpi) and the thickness of the handwriting (10 pixels).

3.3 Directional Roses

The remaining windows are divided into four sub windows of size 16x16 pixels. We compute the autocorrelation function on each of these four windows. The patterns we obtain from this autocorrelation function represent the main directions in the four quadrants of the current window.

We summarize this information in a directional rose of eight directions. Each 16x16 pixels sub window is represented by a signature of eight values resulting in a total of 32 values for the current window. Figure 3 illustrates the sub window extraction process from a selected window taken from a word image from a sample line, and figure 4 shows their respective directional roses. The length of a direction is obtained by summation of the grey levels of the autocorrelation function in this direction.

To keep the most discriminative information, we only keep the relative variations of the different directions above the least represented direction, which is then set to 0. The salient direction is then normalized to 1, to reduce the influence of the dynamic of the original image. Moreover, the use of the autocorrelation function for the local signature reduces the influence of noise or degradations because it is the main structure of the sub image that influences the result. We will give more details on this specific point later in this paper.

As shown in figure 4, the directional rose highlights the dominant orientation in each sub window; the salient direction is quite obvious. This description is only based



Fig. 3. Illustration of the sub windows extraction process and their corresponding autocorrelation functions

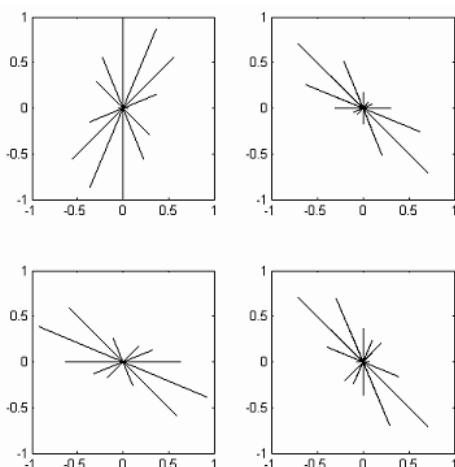


Fig. 4. Directional roses extracted from the four sub windows

on directional information. Even if this directional information is scale invariant, we do not keep this property in our signature because of the decomposition in four windows that have a defined size. However, a normalization step can be performed as a pre-processing to fit the size of the letters in the guide lines of figure 2. Reduction of size introduces no loss of accuracy in the tests we perform because the size we defined for the letters and thus for the words contains enough information. Increasing the size will lead to blurred images. This specific case is discussed latter in this paper.

3.4 The Matching Algorithm

Once the signatures of all the selected windows of each part of the query word image are extracted, the search begins in order to spot all their occurrences. They are compared to all those of the test page, the most similar ones are detected, and the region with most agglomeration of sub windows similar to those of the query word image is considered as a possible match. However, a decision based only on this assumption requires some computation time due to the great number of comparisons to perform. Moreover, as the number of lines and thus windows increases, this first simple matching is not very reliable due to superposition of matching candidates in all three parts of the line resulting in possible yet incorrect matches.

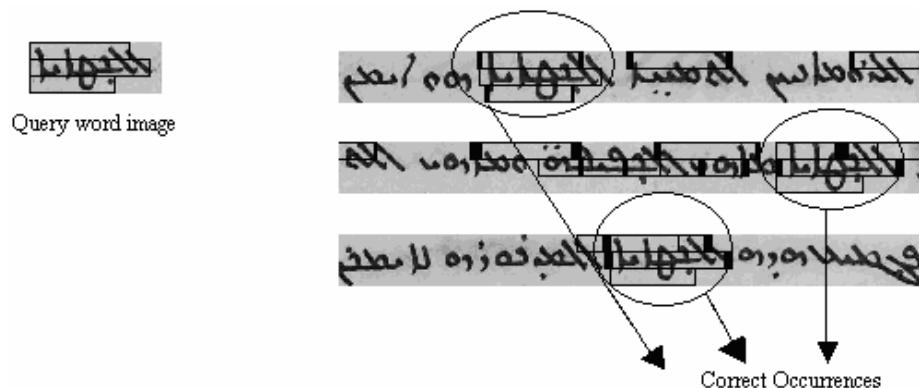


Fig. 5. Detection of regions of interest

In order to surpass the confusion problem, we proceeded by a pre detection of regions of interest where possible occurrences may be located. This was done by studying the movement of the gravity centers of the selected windows in each part of the query word image. The positions of the gravity centers are plotted, and a search is conducted to find portions of the plot that are similar to the query plot based on a minimum Euclidean distance. As a result, confusion is removed as shown in figure 6.

Normalized saliences coefficients are extracted from the query word image and are compared to the ones extracted from the regions of interest. The matching is based on a minimum Euclidean distance, and the correct occurrences are those which have the largest number of matched coefficients in the three superposed parts.

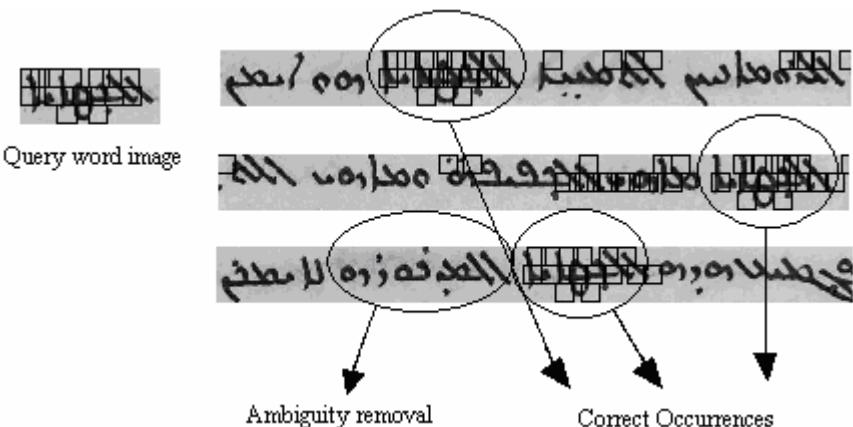


Fig. 6. Ambiguity removal

4 Results

The results we obtain using the combination of these two indicators (directional signature of the windows and pre selected area of interest using gravity center motions) are really promising because, on the tested pages so far, we always find the two or three occurrences of the query word in the top first best matches.

It is difficult to present at the moment, a quantitative evaluation of our method because we do not have yet a real ground truth on our documents. It is not that easy to build it because we have to search manually for words and their occurrences in the documents. We only did this work for some words yet and as mentioned earlier, the results are correct in every tested case so far. A more complete set of test is in progress to give more quantitative results.

Moreover, we tested our indicators in different situations and especially situations involving degraded documents. These degraded versions correspond to the two most common degradations to digitized manuscripts. The first is excessive and lossy compression, and the second is poor resolution.

4.1 Excessive Compression and Poor Resolution

Librarians and book keepers have a tendency to over compress the manuscripts images. In most cases, lossy JPEG compression is chosen. Since the degradations resulting from this compression are irreversible, many people proceed by a restoration phase that consists of a smoothing of the artifacts, sometimes even an attempt to recover the dissolved portions of the texts usually by morphological approaches as used in [16] or by active contours as attempted in [15] and [17], as a result they fall most of the time in a paradigm which is “restore to recognize and recognize to restore”. This is why many approaches fail in front of the degradations introduced by excessive JPEG compression.

The poor resolution type of degradations is introduced either by having manuscripts with large pages, and in an attempt to fit them into a standard size page, librarians and

book keepers tend to reduce the resolution of the digitizer, or just like in the preceding degradation, it is done only by fear of not having enough storage capacity for all the documents. We imitated this degradation by downscaling a test image and afterwards rescaling it back to its original resolution using cubic interpolation.

4.2 Results on Degraded Images

Figure 7 illustrates the ambiguity removal and the ability of finding all occurrences of the query word image. Just as the case for the previously degraded test image, the

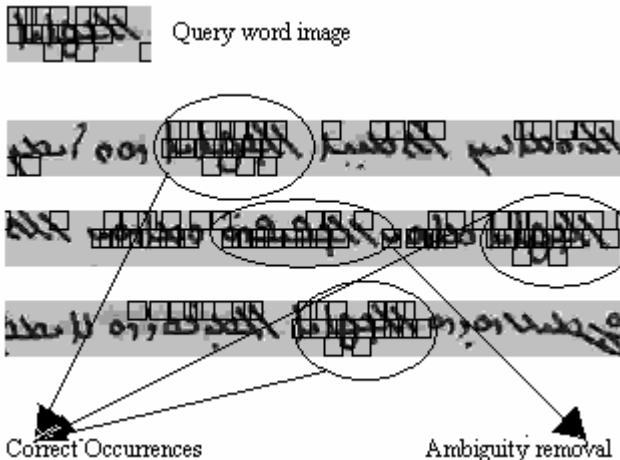


Fig. 7. Ambiguity removal on over-compressed image after interest regions extraction

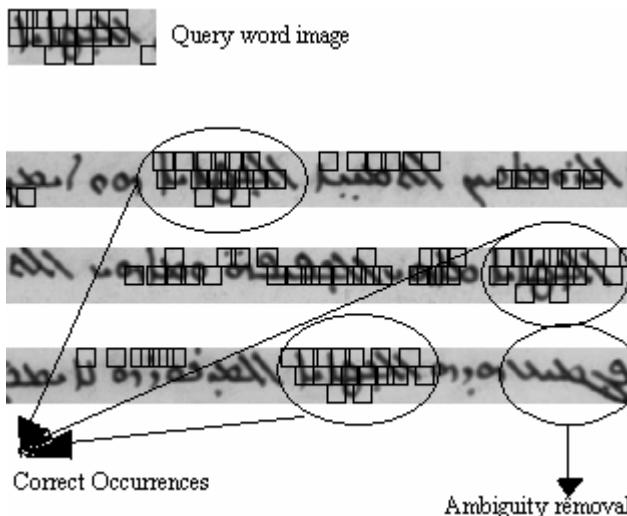


Fig. 8. Ambiguity removal

results for the rescaled test image support the robustness of the algorithm in the face of another type of degradation. Figure 8 reveals the ambiguity removal and the ability to find all occurrences of the query word image after the regions of interest detection.

The results prove the effectiveness of these features and their consistency in finding all occurrences of the query word image within reasonable iterations and processing time. The selective sliding windows, the region of interest detection, the feature extraction, and the matching algorithm were all performed exactly the same as in the word spotting in the original test image which we take now as a reference for the performance comparison

6 Conclusion

In this paper we presented a word spotting algorithm to assist the indexing of Syriac manuscripts. Our method does not require any prior information for the spotting process. It is also fully independent from any word to character segmentation algorithm. Moreover, the way the signature is computed leads to a less sensitivity to noise and degradations that are really common on that type of documents such as an excessive JPEG compression and a rescaling for low resolution documents images. An extension of this method could be used as a basis for a classification algorithm for the automatic separation of the three calligraphies.

References

- [1] Balasubramanian, A., Meshesha, M., Jawahar, C.V.: Retrieval from document image collections. In: Bunke, H., Spitz, A.L. (eds.) DAS 2006. LNCS, vol. 3872, pp. 1–12. Springer, Heidelberg (2006)
- [2] Clocksin, W.F., Fernando, P.P.J.: Towards automatic transcription of Syriac handwriting. In: IEEE Proceedings of the 12th International Conference on Image Analysis and Processing (ICIAP 2003), Mantova, Italy, September 2003, pp. 664–669 (2003)
- [3] Clocksin, W.F.: Handwritten Syriac character recognition using order structure invariance. In: IEEE Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), Cambridge, UK, pp. 562–565 (August 2004)
- [4] Eglin, V., Bres, S., Rivero, C.: Hermite and Gabor transforms for noise reduction and handwriting classification in ancient manuscripts. International Journal on Document Analysis and Recognition (IJDAR 2007) 9, 101–122 (2007)
- [5] Gatos, B., Pratikakis, I., Perantonis, S.J.: An adaptative binarization technique for low quality historical documents. In: Marinai, S., Dengel, A. (eds.) DAS 2004. LNCS, vol. 3163, pp. 102–113. Springer, Heidelberg (2004)
- [6] Gatos, B., Pratikakis, I., Perantonis, S.J.: Adaptive degraded document image binarization., Pattern Recognition. The Journal of the Pattern Recognition Society 39, 317–327 (2006)
- [7] Leydier, Y., Lebourgeois, F., Emptoz, H.: Text search for medieval manuscript images, Pattern Recognition. The Journal of the Pattern Recognition Society 40, 3552–3567 (2007)
- [8] Manmatha, R., Rothfeder, J.L.: A scale space approach for automatically segmenting words from historical handwritten documents. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI 2005) 27, 1212–1225 (2005)

- [9] Nicolas, S., Paquet, T., Heutte, L.: Extraction de la structure de documents manuscrits complexes à l'aide de champs Markoviens. In: Actes du 9ème Colloque International Francophone sur l'Ecrit et le Document (CIFED 2006), pp. 124–129 (September 2006)
- [10] Rath, T.M., Manmatha, R.: Word spotting for historical documents. International Journal on Document Analysis and Recognition (IJDAR 2007) 9, 139–152 (2007)
- [11] Terasawa, K., Nagasaki, T., Kawashima, T.: Eigenspace method for text retrieval in historical document images. In: IEEE Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR 2005), Seoul, Korea, August 2005, pp. 437–441 (2005)
- [12] Terasawa, K., Tanaka, Y.: Locality sensitive pseudo-code for document images. In: IEEE Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Brazil, September 2007, pp. 73–77 (2007)
- [13] Vinciarelli, A., Bengio, S., Bunke, H.: Offline recognition of unconstrained handwritten texts using hmms and statistical language models. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI 2004) 26, 709–720 (2004)
- [14] Weihua, H., Tan, C.L., Sung, S.Y., Xu, Y.: Word shape recognition for image-based document retrieval. In: IEEE Proceedings of the International Conference on Image Processing (ICIP 2001), Thessaloniki, Greece, October 2001, pp. 1114–1117 (2001)
- [15] Allier, B., Emptoz, H.: Degraded character image restoration using active contours: A first approach. In: Proceedings of the ACM Symposium on Document Engineering, Virginia, USA, pp. 142–148 (2002)
- [16] Zheng, Q.J., Kanungo, T.: Morphological degradation models and their use in document image restoration. In: IEEE Proceedings of the International Conference on Image Processing (ICIP 2001), Thessaloniki, Greece, October 2001, pp. 193–196 (2001)
- [17] Allier, B., Bali, N., Emptoz, H.: Automatic accurate broken character restoration for patrimonial documents. International Journal on Document Analysis and Recognition (IJDAR 2006) 8, 246–261 (2006)

A Procedure for Efficient Generation of $1/f^\beta$ Noise Sequences

Youcef Ferdi¹ and Abdelmalik Taleb-Ahmed²

¹ LRES Laboratory, Department of Electrotechnics, University of Skikda
BP.26, Route d'El Hadaïek, Skikda, 21000, Algeria

² LAMIH UMR CNRS 8530 Laboratory, University of Valenciennes and Hainaut
Cambresis, Le Mont Houy, 59313 Valenciennes, France
youcef.ferdi@gmail.com, taleb@univ-valenciennes.fr

Abstract. This paper presents a simple, efficient and fast procedure for generation of $1/f^\beta$ noise sequences. The proposed procedure is based on the impulse invariance method applied to the impulse response of the ideal fractional order integrator whose order $\alpha = \beta/2$ is between 0 and 1. First, an optimal value for the initial value of the impulse response is obtained by minimizing a least squares error criterion and then any of the well-established signal modeling techniques can be employed for the parameterization of the discrete impulse response by pole-zero models. For a given model order, the approximation accuracy depends on the signal modeling technique used. An illustrative example is presented to demonstrate the effectiveness of the method.

Keywords: 1/f-noise, fractional order integration, impulse invariance method, pole-zero model, signal modeling.

1 Introduction

The $1/f^\beta$ noise belongs to a class of random processes whose Power Spectral Density (PSD) follows a power law with a real exponent β between 0 and 2 [1]. Such processes have been observed in various areas [2]. For the purpose of simulation studies, many algorithms have been developed to synthesize sample paths of random processes having power spectrum (at least over several decades of frequency) exhibiting $1/f$ behavior [3-6]. This paper presents a new procedure for generating $1/f^\beta$ noise sequences by filtering Gaussian white noise through a discrete-time fractional order integrator designed using impulse invariance method (excepting the first sample) and signal modeling techniques. From system theory viewpoint, fractional order integrals are computed as outputs of linear invariant systems described by the irrational transfer function $1/s^\alpha$. The case $0 < \alpha < 1$ is of special interest in the context of digital generation of $1/f$ -noise. Unfortunately, the application of fractional integral to the generation of $1/f^\beta$ noise sequences gives rise to two problems: the resulting sampled impulse response has an infinite number of

terms, and its first sample is infinite too. This paper suggests using minimization of a least squares error criterion to obtain the first sample and then applying any of the well-established signal modeling techniques such as Padé, Prony, Shanks methods [7] or Steiglitz-Mcbride technique [8] to find the parameters of a rational approximation model which can be easily and efficiently implemented as a recursive filter. In contrast to previous closely related methods, the proposed approach does not require the use of power series expansion [6, 9], continued fraction expansion [10] or truncation technique [11].

2 Background

Let $w(t)$ be a continuous-time signal. According to Riemann-Liouville definition, the α^{th} -order fractional integral of $w(t)$ is given by [12]

$$x(t) = \int_0^t \frac{(t-\tau)^{\alpha-1}}{\Gamma(\alpha)} w(\tau) d\tau, \quad \alpha > 0. \quad (1)$$

The function $\Gamma(\cdot)$ is the gamma function. In terms of systems, the expression (1) represents a convolution integral which describes the input-output relationship between the input $w(t)$, the output $x(t)$ and the impulse response

$$h_{c\alpha}(t) = \frac{t^{\alpha-1}}{\Gamma(\alpha)} u(t). \quad (2)$$

The signal $u(t)$ is the unit-step function defined as

$$u(t) = \begin{cases} 1, & t \geq 0 \\ 0, & t < 0 \end{cases}. \quad (3)$$

The transfer function of the ideal fractional integrator, i.e., the Laplace transform of $h_\alpha(t)$, is given by

$$H_{c\alpha}(s) = \frac{1}{s^\alpha}. \quad (4)$$

The frequency response can be obtained from (4) by the substitution $s = j2\pi f$ where f is the frequency in Hz.

3 Proposed Method

The analogue impulse response (2) is sampled with the sampling period T_s to yield the discrete impulse response [13]

$$h_\alpha(n) = T_s h_{c\alpha}(nT_s) = \frac{T_s^\alpha n^{\alpha-1}}{\Gamma(\alpha)}, \quad n = 0, 1, 2, \dots. \quad (5)$$

The discrete equivalent of (1) is then given by the discrete convolution

$$x(n) = \sum_{i=0}^{\infty} h_{\alpha}(i)w(n-i) . \quad (6)$$

This is expressed in the z-domain as

$$X(z) = H_{\alpha}(z)W(z) . \quad (7)$$

where $H_{\alpha}(z)$, the z-transform of $h_{\alpha}(n)$, is the transfer function of the discrete time fractional order integrator which performs the digital computation of (1). Applying the impulse invariance method gives rise to two problems here because the first sample of $h_{\alpha}(n)$ is infinite for $0 < \alpha < 1$, and the number of its samples is infinite too.

Selecting the first sample of $h_{\alpha}(n)$

An attempt was made initially to select the initial value using the initial value theorem which states that

$$h_{\alpha}(0) = \lim_{z \rightarrow \infty} H_{\alpha}(z) . \quad (8)$$

Although the transfer function $H_{\alpha}(z)$ is unknown, it can approximately be determined by considering its relationship with (4) in the frequency range $0 \leq f \leq 1/2T_s$ [13]

$$H_{\alpha}(e^{j2\pi f T_s}) \approx \frac{1}{T_s} H_{c\alpha}(j2\pi f) . \quad (9)$$

Such approximation can be achieved by replacing s in (4) with an appropriate s-to-z transform such those of Euler, Tustin, Al-Alaoui's [14] or Tseng [15] given by $1/s \approx T_s/(1-z^{-1})$, $1/s \approx (T_s/2)(1+z^{-1})/(1-z^{-1})$, $1/s \approx (7T_s/8)(1+z^{-1}/7)/(1-z^{-1})$ and $1/s \approx (T_s/6)(7+16z^{-1}+z^{-2})/(3-2z^{-1}-z^{-2})$, respectively. From the above s-to-z transforms, the normalized initial value $h_{\alpha}(0)/Ts^{\alpha}$ has the following form

$$\frac{h_{\alpha}(0)}{Ts^{\alpha}} = C^{\alpha} . \quad (10)$$

while C is a positive real constant. The problem now is how to select the s-to-z transform which yields the best approximation of the desired magnitude response of (4). To solve this problem, we propose a least squares magnitude approximation. Note that according to (13) we are interested here in magnitude response only. The square error for magnitude approximation is defined as [16]

$$E(C) = \sum_{k=1}^N \left| H(e^{j2\pi f_k T_s}) - (2\pi f_k)^{-\alpha} \right|^2 . \quad (11)$$

where $f_k = k \frac{1}{2NT_s}$, $k = 1, 2, \dots, N$ are the frequency samples taken in the range

$0 < f \leq 1/2T_s$ and $H(e^{j2\pi f T_s})$ is the rational approximation defined in (12) whose coefficients a_i and b_i are estimated using signal modeling techniques. It is clear that the optimal value of C which minimizes the error criterion (11) does not necessarily derive from known s-to-z transforms; the latter served only to gain insight into the problem of selecting the initial value $h_{\alpha}(0)$. In figure 1 is plotted the square error as a

function of the normalized initial value C^α for $\alpha = 0.5$, $T_s = 0.001$, $L = 1000$ and $p = q = 5$ using MATLAB's m-file prony. The computation of the square error has been made by varying one of the parameters α , T_s or (p, q) while holding the others constant. In all the tests carried out, with $0.1 \leq C \leq 4$, the square error presents a unique minimum corresponding to a normalized optimal initial value located near 1. The value $C_0^\alpha = 1$ will then be taken as an initial guess for the optimization function.

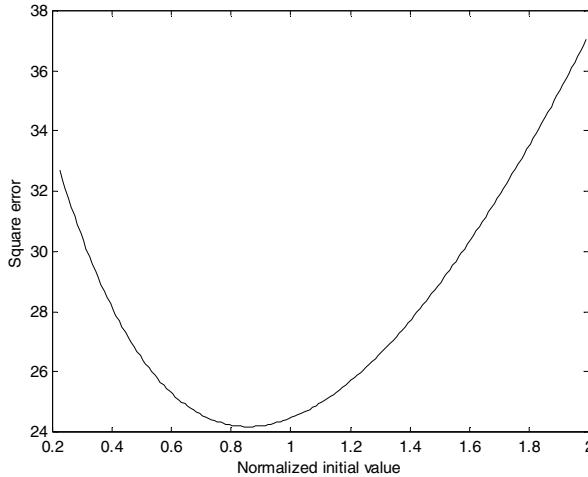


Fig. 1. Square error $T_s = 0.001$, $L = 1000$ and $p = q = 5$. $\alpha = 0.4$

Approximating $H_\alpha(z)$

There are two approaches that can be used for the approximation of $H_\alpha(z)$: 1) direct truncation of (5) to N terms and, 2) signal modeling techniques using the first L samples of (5). The former suffers from the Gibbs phenomenon and requires lengthy FIR filter as the decay of $h_\alpha(n)$ is slow. The latter can lead to IIR filter approximation whose transfer functions possess the desired properties, namely, stability, minimum-phase, and zeros and poles distributed in alternating fashion on the real axis. Given the fractional order α , the sampling period T_s , the first L samples of (5) and the degrees p and q of the approximating rational transfer function (12)

$$H(z) = \frac{X(z)}{W(z)} = \sum_{i=0}^q b_i z^{-i} \left/ \left(1 + \sum_{i=1}^p a_i z^{-i} \right) \right.. \quad (12)$$

The unknown coefficients, a_i and b_i , are to be determined such that $H(z)$ best approximates $H_\alpha(z)$ in some sense. The generated sequences $x(n)$ will have PSDs proportional to the frequency response squared magnitude, that is

$$S_{xx}(f) = \sigma_w^2 |H_\alpha(e^{j2\pi f T_s})|^2. \quad (13)$$

where σ_w^2 is the variance of the Gaussian white noise $w(n)$. These PSDs approximate, over a finite range of frequencies, the perfect PSD of the $1/f^\beta$ noise.

Determining the optimal initial value

The optimization of the error criterion (11) is a one dimensional unconstrained optimization problem. A suitable search method for this type of problems is the simplex method of Nelder and Mead which requires function evaluations only. This method is implemented in MATLAB by the function *fminsearch*. The MATLAB function to be provided to *fminsearch* is the following

```
function E = optinivalue(C,Ts,L,p,q,r,f);
%function to be minimized to find the
%optimal value of the first sample
% r is the fractional order
n = 1:L;
h = [(C*Ts)^(r), (Ts^(r))*(n.^(r-1))/gamma(r)];
[b,a]=prony(h,q,p); % or stmcb
H = freqz(b,a,f,1/Ts);
Hid = (2*pi*f).^( -r); %ideal frequency %response
E = sum((Hid-abs(H)).^2);
```

Table 1. Approximate transfer functions to $1/s^\alpha$

$\alpha = 0.4$, prony	$\frac{0.05936 - 0.1588z^{-1} + 0.1483z^{-2} - 0.0545z^{-3} + 0.005409z^{-4} + 0.0003039z^{-5}}{1 - 3.155z^{-1} + 3.694z^{-2} - 1.939z^{-3} + 0.4258z^{-4} - 0.02601z^{-5}}$
$\alpha = 0.8$, prony	$\frac{0.003897 - 0.009623z^{-1} + 0.007922z^{-2} - 0.002234z^{-3} + 4.021e-006z^{-4} + 3.955e-005z^{-5}}{1 - 3.347z^{-1} + 4.205z^{-2} - 2.411z^{-3} + 0.5966z^{-4} - 0.04417z^{-5}}$
$\alpha = 0.4$, stmcb	$\frac{0.05936 - 0.2145z^{-1} + 0.2909z^{-2} - 0.1774z^{-3} + 0.04327z^{-4} - 0.001651z^{-5}}{1 - 4.092z^{-1} + 6.542z^{-2} - 5.067z^{-3} + 1.876z^{-4} - 0.2589z^{-5}}$
$\alpha = 0.8$, Stmcb	$\frac{0.003683 - 0.01264z^{-1} + 0.01578z^{-2} - 0.008149z^{-3} + 0.001118z^{-4} + 0.0002107z^{-5}}{1 - 4.36z^{-1} + 7.514z^{-2} - 6.38z^{-3} + 2.659z^{-4} - 0.4324z^{-5}}$

The parameters α and T_s are application-dependent while L , p , q and the signal modeling techniques are to be chosen by the designer. Taking an initial value $C_0 = 1$, the transfer functions reported in Table I were obtained with $\alpha = 0.4$ and $\alpha = 0.8$, $T_s = 0.001$, $L = 1000$, $N = 2^{16}$ and $p = q = 5$ using Prony and Steiglitz-McBride methods. The latter is implemented in the MATLAB's m-file *stmcb*. The checking of the pole-zero maps show that the poles and zeros of the transfer functions in Table I are inside the unit circle and are distributed in alternating fashion on the real axis and, hence, they correspond to stable minimum phase systems. The magnitude responses are plotted in figure 2. This figure shows that the magnitude response is better approximated at low frequency using Steiglitz-McBride's method than Prony's method. Experiments have shown that increasing the model order results in slight improvement in the quality of the magnitude approximation. For example, the square error (11) was found to be equal respectively to 53.9906, 42.0917 and 32.5782 for $p = q = 4, 5$ and 6 with $\alpha = 0.8$ and the same values for the others parameters using Steiglitz-McBride's method.

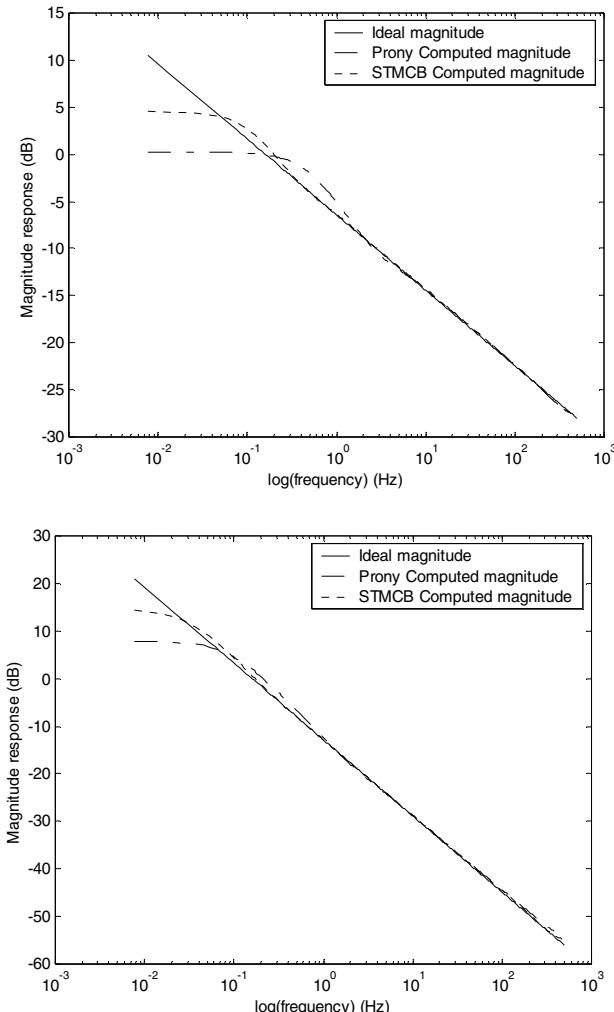


Fig. 2. Magnitude frequency response for $T_s = 0.001$, $L = 1000$ and $p = q = 5$. Top: $\alpha = 0.4$, bottom: $\alpha = 0.8$.

4 Illustrative Example

The proposed procedure was used to generate 100 $1/f^\beta$ noise sequences of 2^{10} samples by filtering zero-mean, unit variance Gaussian white noise through discrete time fractional order integrator whose transfer functions are given in Table I. The periodogram of each sequence was computed by the use of FFT technique. Estimates of the PSDs were obtained as averages over the 100 periodograms and plotted in figure 3 along with the frequency response squared magnitude of the filter used for

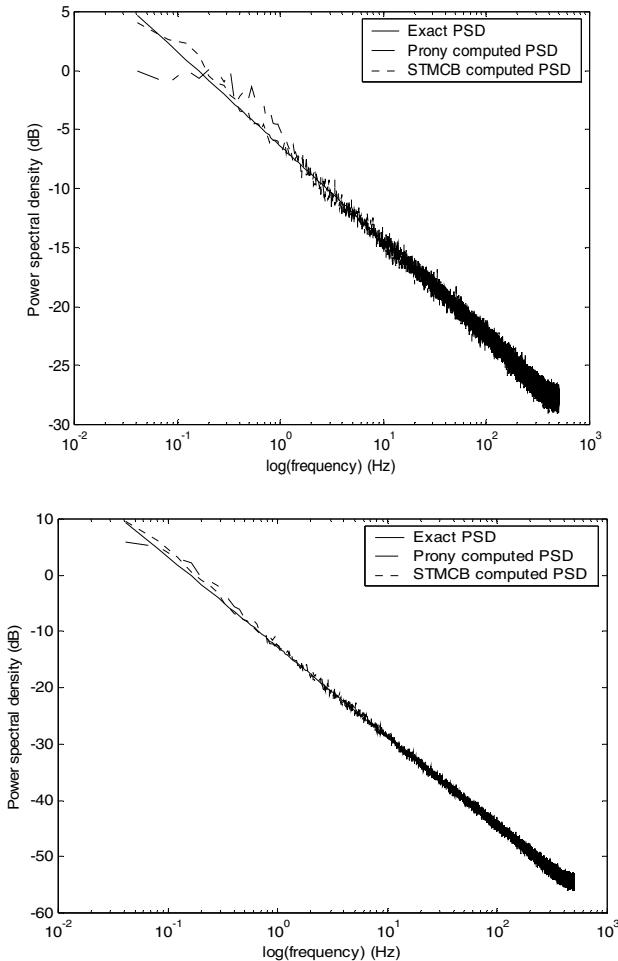


Fig. 3. Average power spectral densities for $T_s = 0.001$, $L = 1000$ and $p = q = 5$. Top: $\alpha = 0.4$, bottom: $\alpha = 0.8$.

generation. As it can be seen from this figure, the PSD estimates of the generated noise sequences best fit the exact PSD over a frequency band of about four decade using Steiglitz-McBride's technique than Prony's method.

5 Conclusion

This paper describes a simple, efficient and fast method for the generation of $1/f^\beta$ noise sequences via impulse invariance method and signal modeling techniques. The first sample of the impulse response is optimally selected and then the resulting sampled one is modeled as the impulse response of a linear invariant system. The accuracy of the approximation is influenced by signal modeling technique used. The

proposed procedure can be easily implemented to design recursive filters for efficient generation of $1/f^\beta$ -noise sequences from white noise sequences. In contrast to other methods, the proposed approach does not require the use of power series expansion, continued fraction expansion or truncation technique.

References

1. Keshner, M.S.: 1/f Noise. Proceedings of the IEEE 70(3), 212–218 (1982)
2. Wornell, G.W.: Signal processing with fractals: A wavelet-based approach. Prentice-Hall, Englewood Cliffs (1995)
3. Flandrin, P.: Wavelet analysis and synthesis of fractional Brownian motion. IEEE Trans. Inform. Theory 38(2), 910–917 (1992)
4. Perrin, E., Harba, R., Jennane, R.: Fast and exact synthesis of 1D fractional Brownian motion. IEEE Signal Proc. Letters 9(11), 382–384 (2002)
5. Guglielmi, M.: $1/F^\alpha$ signal synthesis with precision control. Signal Processing 86, 2548–2553 (2006)
6. Rodriguez, E., Echeverria, J.C., Ramirez, J.A.: $1/F^\alpha$ fractal signal generation from Grünwald-Letnikov formula. Chaos, Solitons and Fractals (in press, 2008)
7. Proakis, J.G., Manolakis, D.G.: Digital Signal Processing. Principles, Algorithms, and Applications. Prentice-Hall international, Englewood Cliffs (1996)
8. Steiglitz, K., McBride, L.E.: A Technique for the Identification of Linear Systems. IEEE Transactions on Automatic Control AC-10, 461–464 (1965)
9. Ferdi, Y.: Computation of Fractional Order Derivative and Integral via Power Series Expansion and Signal Modeling. Nonlinear Dynamics 46(1-2), 1–15 (2006)
10. Chen, Y.Q., Vinagre, B.M., Podlubny, I.: Continued fraction expansion approaches to discretizing fractional order derivatives—an expository review Nonlinear Dynamics 38(1-2), 155–170 (2004)
11. Lee, S., Zhao, W., Narasimha, R., Rao, R.M.: Discrete-time models for statistically self-similar signals. IEEE Transactions on Signal Processing 51(5), 1221–1230 (2003)
12. Oldham, K.B., Spanier, J.: The Fractional Calculus. Academic Press, New York (1974)
13. Oppenheim, A.V., Schafer, R.W.: Digital Signal Processing, pp. 198–203. Prentice-Hall, Englewood Cliffs (1975)
14. Al-Alaoui, M.A.: Novel Digital Integrator and Differentiator. Electronics letters 29(4), 376–378 (1993)
15. Tseng, C.-C.: Digital Integrator Design Using Simpson Rule and Fractional Delay Filter. IEE Proc.-VIS, Image Signal Process 153(1), 79–85 (2006)
16. Steiglitz, K.: Computer-aided design, of recursive digital filters. IEEE Transactions on Audio and Electroacoustics AU-18, 123–129 (1970)

Design and Implementation of an Image CoProcessor

R. Ebrahimi Atani¹, S. Mirzakuchaki¹, and S. Ebrahimi Atani²

¹ EE Department, IUST, Narmak 16846, Tehran, Iran
rebrahimi@iust.ac.ir, M_kuchaki@iust.ac.ir
<http://webpages.iust.ac.ir/rebrahimi/>

² Department of Mathematics, Guilan University,
Faculty of Science, P.O. Box 1914, Rasht, Iran
ebrahimi@gilan.ac.ir

Abstract. This paper presents a novel DA based 2D DCT/DST coprocessor architecture for the synchronous design in a Xilinx FPGA device. A 1.2V, 90nm triple-oxide technology, Virtex-IV FPGA is used for final implementation and maximum operating frequency of 117 MHz is achieved. Using XPower toolbox, the total dynamic power consumption of 393 mW is measured. The paper presents the trade-offs involved in designing the architecture, and the design for performance issues.

1 Introduction

Field programmable gate array (FPGA) digital signal processing (DSP) systems are increasingly employed by designers to address demanding digital communications, video, and imaging applications. Usually, FPGAs are used as VLSI replacement on low volume production or prototyping devices which are to be eventually implemented as application specific ICs (ASIC). The latest DSP applications demand system architecture solutions that address a range of requirements including performance, flexibility and scalability. FPGA coprocessors can offload a DSP processor and efficiently execute computationally intensive blocks of a DSP algorithm due to the inherent parallelism. This is especially attractive for emerging applications such as video and image processing and high speed digital communications where DSP performance requirements are growing at the fastest rates. The designer of DA based DSP algorithms in FPGAs is confronted with more design choices than the DSP programmer faces with the standard DSP microprocessor chips. Indeed, the number formats of the DSP microprocessor are pre-ordained either fixed point 16 or 24 bit, or 32 bit floating point word lengths. The FPGA design choice ranges from 2 bits to as many bits as are required; data word size and coefficient word size may be independently set, and may be modified along the data path all subject to the constraint of a fixed point number format. The choices available to the FPGA designer present both challenges and opportunities. Designs can be optimized in both performance and number of gate and memory resources.

Digital signal processing transforms (DXT) are required in almost all the phases of image and signal processing applications. DXTs are employed for various imaging techniques and spectral analysis of the signals. The discrete cosine transform (DCT) and discrete Sine transforms are two very well known DXTs that are widely used for VLSI digital signal and image processing systems. These two transforms have been used in image and speech processing and are the optimum and fast algorithms for image and data compression applications. DCT forms a key role in many image and video compression standards including JPEG2000 (image compression), ITU H.261 and ITU H.263 (Tele-conferencing), and ISO MPEG1 and MPEG2 (home video) [11]. DCT is also widely considered to be the most effective scheme for transform coding. DST is useful for spectrum analysis, data compression, speech processing, biomedical signal processing and in many other applications.

Several algorithms and implementations have been proposed for the DXTs, spreading from software implementations in DSPs to hardware implementations in ASICs ([1], [2], [3], [7], and [10]). The DXTs are computationally intensive and as such, there is a great demand for high speed, high throughput and short latency computing architectures. Due to the high computation requirements, 2D DCT/DST processor design has been concentrated on small non overlapping blocks (typical 8×8 or 16×16). Many 2D DCT/DST algorithms have been proposed to achieve reduction of the computational complexity and thus increase the operational speed and throughput. The various algorithms and architectures for the 2D DCT can be divided into two categories:

1. Row/Column decomposition methods
2. Non row/column decomposition methods

There are some implementations for the DXT calculations, but two of the more important ones are systolic architecture (SA) and distributed arithmetic (DA). The main goal of the paper is the design of an 8×8 DA-based DCT/DST coprocessor. In this paper we will use DA due to their suitability for FPGA implementation and because they show dramatic improvements and better performance in comparison with the SA technique (in terms of speed and area consumption). We will also exploit row/column decomposition for parallelization and pipelining of the architecture. Paper presents the trade-offs involved in designing the architecture, and the design for performance issues.

This paper is structured as follows: a brief overview of DA is presented in Section 2. 1D DA based DCT/DST core architectures are described in section 3. In section 4, 2D DCT/DST coprocessor design flow is presented. Implementation results are discussed in Section 5, and finally, conclusions are in Section 6.

2 Distributed Arithmetic

Distributed arithmetic is a bit level rearrangement of a multiply accumulate to hide the multiplications. DA plays a key role in embedding DSP functions in FPGA devices because the algorithm is based on Look up table (LUT). DA is

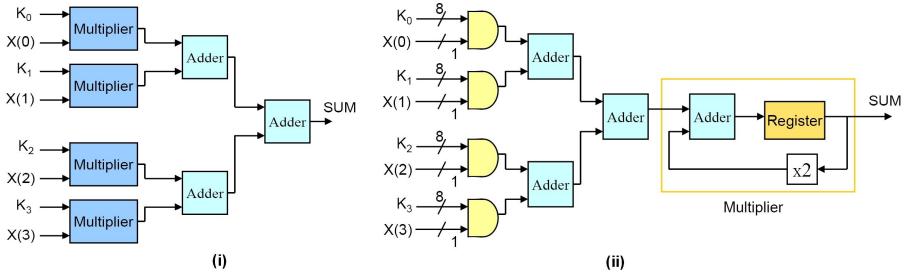


Fig. 1. (i) 4-product MAC using Conventional Arithmetic (ii) 4-product MAC using Serial Distributed Arithmetic

one of the most common techniques where the multiply - accumulate (MAC) is of paramount importance. Its greatest advantage is in using additions instead of multiplications which is desirable since a multiplication consumes much more hardware than an addition. We can implement the 4-product MAC in two ways: one using conventional arithmetic and the other with serial DA as is shown in Fig. 1(i, ii). It is clearly obvious that great reductions on the hardware will be achieved when using DA. So In our implementation we have employed the 4 product MAC using Serial DA. The 4-product MACs, each consists of four multiplication units that are added together to produce the SUM output as illustrated in Fig. 1(i). In this design, the architecture in Fig. 1(ii) is used which produces the same result.

3 1D DA Based DCT/DST Architecture

Since the 1D DCT/DST are the basic elements for the 2D DCT/DST implementation, first the DCT and DST transforms of 1D will be presented. In case of 1D, for any real sample sequence $x(n)$ where $n \in (0, 1, \dots, N - 1)$, the DCT and DST can be defined as:

$$DCT : \begin{cases} C(k) = \sqrt{\frac{2}{N}} \cdot \alpha_k \sum_{n=0}^{N-1} x(n) \cos[(2n+1) \cdot \frac{k \cdot \pi}{2N}] \\ k = 0, 1, \dots, N-1 \text{ where } \alpha_k = \begin{cases} \frac{1}{\sqrt{2}} & k = 0 \\ 1 & \text{otherwise} \end{cases} \end{cases} \quad (1)$$

$$DST : \begin{cases} S(k) = \sqrt{\frac{2}{N}} \cdot \beta_k \sum_{n=0}^{N-1} x(n) \sin[(2n+1) \cdot \frac{k \cdot \pi}{2N}] \\ k = 1, 2, \dots, N \text{ where } \beta_k = \begin{cases} \frac{1}{\sqrt{2}} & k = N \\ 1 & \text{otherwise} \end{cases} \end{cases} \quad (2)$$

The direct implementation of these equations need N^2 multiplications. Because cyclic properties and symmetry of sinusoidal functions, the number of

multiplications can be reduced. According to the definition of DCT in (1), a different representation for the DCT data sequence is shown in (2).

$$C(k) = \sum_{n=0}^{N-1} A_{n,k} \cdot x(n) \quad (3)$$

Where, A is the transform matrix. The real-time computation of DCT requires a lot of calculations because of the large number of multiplications involved, therefore much effort has gone into reducing the total number of multiplications. Various authors have proposed diverse algorithms for this reduction, like matrix decomposition, polynomial transformation or combinations of both. By using matrix decomposition and taking advantage of the coefficient symmetry from Chen et al [6], it can be shown that we can divide the transform matrix A into two smaller matrices, significantly reducing the total number of multiplications to $\frac{N^2}{2}$. Then DCT is obtained easily from (4) and (5):

$$C(k) = \sum_{n=0}^{\frac{N}{2}-1} A_{n,k} \cdot [x(n) + x(N-1-n)] \quad \text{for even } k \quad (4)$$

$$C(k) = \sum_{n=0}^{\frac{N}{2}-1} A_{n,k} \cdot [x(n) - x(N-1-n)] \quad \text{for odd } k \quad (5)$$

In case of an 8-point transform, due to the symmetry, A can be replaced by two matrices which can be computed in parallel, according to (6) and (7):

$$\begin{bmatrix} Y(0) \\ Y(2) \\ Y(4) \\ Y(6) \end{bmatrix} = \begin{bmatrix} 0.5 \cos\left(\frac{\pi}{4}\right) & 0.5 \cos\left(\frac{\pi}{4}\right) & 0.5 \cos\left(\frac{\pi}{4}\right) & 0.5 \cos\left(\frac{\pi}{4}\right) \\ 0.5 \cos\left(\frac{\pi}{8}\right) & 0.5 \cos\left(\frac{3\pi}{8}\right) & -0.5 \cos\left(\frac{3\pi}{8}\right) & -0.5 \cos\left(\frac{\pi}{8}\right) \\ 0.5 \cos\left(\frac{\pi}{4}\right) & -0.5 \cos\left(\frac{\pi}{4}\right) & -0.5 \cos\left(\frac{\pi}{4}\right) & 0.5 \cos\left(\frac{\pi}{4}\right) \\ 0.5 \cos\left(\frac{3\pi}{8}\right) & -0.5 \cos\left(\frac{\pi}{8}\right) & 0.5 \cos\left(\frac{\pi}{8}\right) & -0.5 \cos\left(\frac{3\pi}{8}\right) \end{bmatrix} \begin{bmatrix} x(0) + x(7) \\ x(1) + x(6) \\ x(2) + x(5) \\ x(3) + x(4) \end{bmatrix} \quad (6)$$

$$\begin{bmatrix} Y(1) \\ Y(3) \\ Y(5) \\ Y(7) \end{bmatrix} = \begin{bmatrix} 0.5 \cos\left(\frac{\pi}{16}\right) & 0.5 \cos\left(\frac{3\pi}{16}\right) & 0.5 \cos\left(\frac{5\pi}{16}\right) & 0.5 \cos\left(\frac{7\pi}{16}\right) \\ 0.5 \cos\left(\frac{3\pi}{16}\right) & -0.5 \cos\left(\frac{7\pi}{16}\right) & -0.5 \cos\left(\frac{5\pi}{16}\right) & -0.5 \cos\left(\frac{3\pi}{16}\right) \\ 0.5 \cos\left(\frac{5\pi}{16}\right) & -0.5 \cos\left(\frac{\pi}{16}\right) & -0.5 \cos\left(\frac{7\pi}{16}\right) & 0.5 \cos\left(\frac{5\pi}{16}\right) \\ 0.5 \cos\left(\frac{7\pi}{16}\right) & -0.5 \cos\left(\frac{5\pi}{16}\right) & 0.5 \cos\left(\frac{3\pi}{16}\right) & -0.5 \cos\left(\frac{\pi}{16}\right) \end{bmatrix} \begin{bmatrix} x(0) - x(7) \\ x(1) - x(6) \\ x(2) - x(5) \\ x(3) - x(4) \end{bmatrix} \quad (7)$$

It should be noted that the calculation of DST is similar to DCT with the only distinction being the different coefficient matrices and we are not going into details about it any further. (8) and (9) show the DST coefficients. Fig. 2 illustrates the overall architectures for 8-point 1D DCT and 1D DST. The four sum boxes are the equivalents of the corresponding sums in parenthesis in (6) and (9) and also the four difference boxes are coming from differences in (7) and (8). In the next stage we have 4-product MACs which each one consists of four multiplication units that are added together to produce the SUM output. We can implement the 4-product MAC using serial distributed arithmetic as illustrated

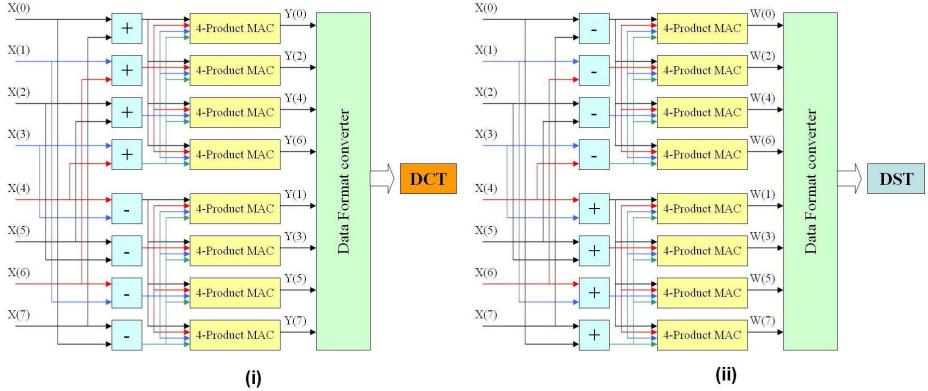


Fig. 2. (i) 8-point 1-D DCT (ii) 8-point 1-D DST

in Fig. 1 (ii). If we examine (6) and (7) again, we can observe that the coefficient matrices are 4×4 and that is why we have used 4-product MACs. Each 4-product MAC calculates the multiplication of two vectors: one row of the coefficient matrix and the other vector consisted of four sum boxes (or difference boxes). The outputs of the MACs are not in the correct order and need reordering. In the final stage there is a Data Format Converter that has the task of re-arranging the outputs of the MACs and delivers the final DCT or DST result.

$$\begin{bmatrix} W(8) \\ W(2) \\ W(4) \\ W(6) \end{bmatrix} = \begin{bmatrix} 0.5 \sin(\frac{\pi}{4}) & 0.5 \sin(\frac{\pi}{4}) & 0.5 \sin(\frac{\pi}{4}) & 0.5 \sin(\frac{\pi}{4}) \\ 0.5 \sin(\frac{\pi}{8}) & 0.5 \sin(\frac{3\pi}{8}) & -0.5 \sin(\frac{3\pi}{8}) & -0.5 \sin(\frac{\pi}{8}) \\ 0.5 \sin(\frac{\pi}{4}) & -0.5 \sin(\frac{\pi}{4}) & -0.5 \sin(\frac{\pi}{4}) & 0.5 \sin(\frac{\pi}{4}) \\ 0.5 \sin(\frac{3\pi}{8}) & -0.5 \sin(\frac{\pi}{8}) & 0.5 \sin(\frac{\pi}{8}) & -0.5 \sin(\frac{3\pi}{8}) \end{bmatrix} \begin{bmatrix} x(0) - x(7) \\ x(1) - x(6) \\ x(2) - x(5) \\ x(3) - x(4) \end{bmatrix} \quad (8)$$

$$\begin{bmatrix} W(1) \\ W(3) \\ W(5) \\ W(7) \end{bmatrix} = \begin{bmatrix} 0.5 \sin(\frac{\pi}{16}) & 0.5 \sin(\frac{3\pi}{16}) & 0.5 \sin(\frac{5\pi}{16}) & 0.5 \sin(\frac{7\pi}{16}) \\ 0.5 \sin(\frac{3\pi}{16}) & -0.5 \sin(\frac{7\pi}{16}) & -0.5 \sin(\frac{\pi}{16}) & -0.5 \sin(\frac{5\pi}{16}) \\ 0.5 \sin(\frac{5\pi}{16}) & -0.5 \sin(\frac{1\pi}{16}) & -0.5 \sin(\frac{3\pi}{16}) & 0.5 \sin(\frac{7\pi}{16}) \\ 0.5 \sin(\frac{7\pi}{16}) & -0.5 \sin(\frac{5\pi}{16}) & 0.5 \sin(\frac{3\pi}{16}) & -0.5 \sin(\frac{\pi}{16}) \end{bmatrix} \begin{bmatrix} x(0) + x(7) \\ x(1) + x(6) \\ x(2) + x(5) \\ x(3) + x(4) \end{bmatrix} \quad (9)$$

The reduction in hardware in distributed arithmetic is evident from the previous figures. We have used ROMs to store the coefficients needed, speeding up the transform calculation by a large degree. But using ROMs has the disadvantage of adding to the hardware cost, which is undesirable. It can be ameliorated by using the CORDIC algorithm to calculate the coefficients but in real-time applications such as video and image processing it is not feasible and the use of the CORDIC algorithm is left for special CORDIC processors where speed is not vital.

4 2D DCT/DST Architecture and Design Flow

In this section the design flow of the image coprocessor is discussed. The equations for the $N \times N$ point 2D DXT of an input sequence $\{x(m, n) : m = n = 0, 1, \dots, N - 1\}$ can be shown as:

$$2DDCT : \begin{cases} C(k, l) = \sqrt{\frac{2}{N}} \cdot \alpha_k \alpha_l \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} x(m, n) \cos[(2m + 1) \cdot \frac{k \cdot \pi}{2N}] \cdot \cos[(2n + 1) \cdot \frac{l \cdot \pi}{2N}] \\ k, l = 0, 1, \dots, N - 1 \text{ where } \alpha_k = \alpha_l = \begin{cases} \frac{1}{\sqrt{2}} & k = l = 0 \\ 1 & \text{otherwise} \end{cases} \end{cases} \quad (10)$$

$$2DDST : \begin{cases} S(k, l) = \sqrt{\frac{2}{N}} \cdot \beta_k \beta_l \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} x(m, n) \sin[(2m + 1) \cdot \frac{k \cdot \pi}{2N}] \cdot \sin[(2n + 1) \cdot \frac{l \cdot \pi}{2N}] \\ k, l = 0, 1, \dots, N - 1 \text{ where } \beta_k = \beta_l = \begin{cases} \frac{1}{\sqrt{2}} & k = l = 0 \\ 1 & \text{otherwise} \end{cases} \end{cases} \quad (11)$$

It is easy to see that, direct implementation of (10) or (11) needs N^4 multiplications. But in real time processing with hardware requirements and energy consumption considerations, this is infeasible. Reordering (10) and (11), it can be broken down in two groups of N 1D-DXT, which is equivalent to processing a data block by rows, followed by a column processing or vice versa. This is the most widely utilized method due to its simplicity, regular form and easy implementation. By using matrix notation, the $N \times N$ point 2D DXT of (10) or (11) can be expressed as a matrix vector computation:

$$[T] = [M_{N \times N}].[x] \quad (12)$$

Where M represents the DXT coefficient matrix. By applying row column decomposition, the algorithm can be rewritten using two 1D DXTs and a matrix transpose, as shown in (13):

$$[T] = [Q_{1 \times N}].[x].[Q_{N \times 1}]^t \quad (13)$$

This method needs a transposition memory between the two 1D DXT blocks. If the timing specifications of the application permit it, only one N 1D DXT block can be used for both row and column processing. It is necessary to add a multiplexer for the selection of the input to the DXTs, but the N 1D DXT area is saved. The drawback of this method is the limited bandwidth of transposition memory. In the design of image coprocessor two 8 point 1D DXT is used and 8×8 2D DXT is applied to the input image.

4.1 Design Strategy

In this section we will describe the design strategy of the 2D image coprocessor. The transformation core has simple input interface. Design is synchronous and whole architecture use two 1D DCT and two 1D DST blocks to produce 2D

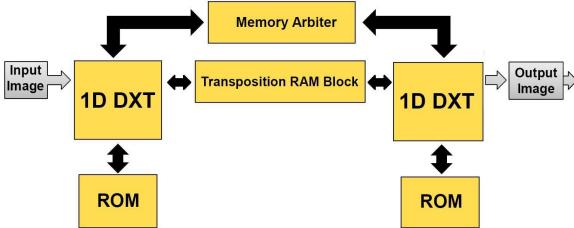


Fig. 3. Architecture for 2D DXT coprocessor

trasforms. A multiplexer is used to distinguish between DST and DCT transforms. The block diagram of the whole architecture is shown in Fig. 3. Two 1D DXT units are connected through transposition matrix memory. Latency between first latched input data and first DXT transformed output is 85 clock cycles. Design is internally pipelined, when the pipeline is full 64 point input data is transformed in 64 clock cycles to 2D DXT values.

Core uses double buffered RAM for storing intermediate product results after first DXT stage for maximized performance. This way both 1D DXT units can work in parallel effectively creating dual stage global pipeline. 2D DXT core takes 8 bit input data and produces 12 bit output using 12 bit DXT matrix coefficients. This may be enhanced to be configurable in the future. Transposition RAM is double buffered, that is when second stage of DXT reads out data from transposition memory 1, first DXT stage can populate second transposition memory with new data. This enables creation of dual stage global pipeline where every stage consists of 1D DXT and transposition memory. 1D DXT units are not internally pipelined; they use parallel DA to compute DXT values. Because of parallel DA they need considerable amount of ROM memories to compute one DXT values in single clock cycle. Design based on DA does not use any multipliers for computing MAC, instead it stores precomputed MAC results in ROM memory and grab them as needed. ROMs can be either asynchronous or synchronous. A memory arbiter between DXT stages is essentially needed. Each stage can request memory buffer for read/write and basing on availability one of two RAM buffers is granted. Every RAM has 10 bit width data and 64 memory cells. On chip RAMs used are dual port synchronous memory with one clock cycle delay. These RAMs are currently 64 words of 14 bits size and there are two of them to create ping pong buffers for improved throughput.

4.2 Testbench

In order to test the applicability of the architecture, a ModelSim based testbench is created. The testbench takes Lena image as input (using Matlab image to text function). Image coprocessor transforms the Lena image to 2D DCT or 2D DST matrices and then by ModelSIM and a behavioral inverse transform reconstructed image is obtained. To compare the quality of reconstruction,



Fig. 4. (i) Original image (ii) 2D-DCT reconstructed image (iii) 2D-DST reconstructed image

the peak signal to noise ratio (PSNR) is computed between original and reconstructed images. In case of (512×512) pixel size Lena image, PSNR of 2D-DCT and 2D-DST reconstructed images was 45.5 dB and 45 dB respectively. The obtained PSNRs are above the required margin of 40 dB. The introduced errors are because of fixed point arithmetic operations. The original Lena image and reconstructed 2D-DCT/DST images are shown in Fig. 4.

5 Implementation Results

The whole architecture including the computation block, data path, and control unit is modeled at register transfer level in VHDL, simulated and tested by a testbench using Modelsim and Matlab. The target device for implementation is a 90 nm VertexIV Xilinx FPGA. The hardware description of this architecture was synthesized using ISE Xilinx 8.1 and mapped on the VertexIV (4VSX35FF668-12). The timing analysis showed that, the worst delay time is less than 8.5 ns which set up the maximum clock of the system to 117 MHz. The architecture's set up slack is 1.546 nano seconds. 56% of data arrival time is because of cell

Table 1. Device Utilization for Vertex-IV (4VSX35FF668-12)

Resource	Used	Available	Utilization
IOs	33	448	7.37%
Global Buffers	1	32	3.13%
Function Generators	1012	30720	3.3%
CLB Slices	563	15360	3.7%
Dffs or Latches	579	31616	1.83%
Block RAMs	42	192	21.88%
DSP48s	0	192	0.00%

delay and the remaining 44% is due to net delay. Power consumption of the architecture has been estimated using XPower toolbox of ISE with an input stimuli showing the clocking and toggling rates of the architecture's components. The power summary showed that the total dynamic power dissipation of this implementation is 393 mW. The resource utilization has been demonstrated in table 1. Less than 7% of the LUTs of the FPGA chip has been used.

6 Summary and Conclusions

This paper has proposed an efficient mapping on FPGA of an image coprocessor. A fully parallel architecture based on row/column decomposition is proposed for the computation of the 2D DA based 8×8 DCT/DST on a Vertex IV Xilinx FPGA. The system is highly modular and utilizes a highly parallel structure to achieve high speed performance. It uses two identical units for the computation of the row and column transforms. We have exploited DA in order to achieve surface reduction and precision amelioration compared to conventional algorithm.

Acknowledgment

Reza Ebrahimi Atani receives partial funding through Iran Telecommunication Research Center (<http://www.itrc.ac.ir>).

References

1. Amiri, M., Ebrahimi Atani, R., Mirzakuchaki, S., Mahdavi, M.: Design and Implementation of 50 MHz DXT Coprocessor. In: 10th Euromicro Conference on Digital System Design, Architectures, Methods and Tools, August 29-31, 2007, pp. 43–50. IEEE press, Lubeck (2007)
2. EL-Bannai, A., EL-Fattah, A., Fakhr, W.: An efficient implementation of the 1D DCT using FPGA technology. In: 11th IEEE International Conference and Workshop on the Engineering of Computer-Based Systems, May 24-27, 2004, vol. 27, pp. 356–360 (2004)
3. Ebrahimi Atani, R., Mirzakuchaki, S., Samii, F., Nasrollahzadeh, M.: Design and Implementation of a 157 MHz DA-Based DXT Coprocessor. In: 4th IEEGCC conference, Manama, Kingdom of Bahrain, November 11-14 (2007)
4. Ahmed, N., Natarajan, T., Rao, K.R.: Discrete cosine transform. IEEE Transactions on Computers C-23(1), 90–93 (1974)
5. Wu, H.R., Paoloni, F.J.: A two-dimensional fast cosine transform algorithm based on Hou's approach. IEEE Transactions on Signal Processing 39(2), 544–546 (1991)
6. Chen, W.H., Smith, C.H., Fralick, S.C.: A fast computational algorithm for the discrete Cosine transform. IEEE Transactions on Communications 25(9), 1004–1009 (1977)
7. Ghosh, S., Venigalla, S., Bayoumi, M.: Design and Implementation of a 2D-DCT Architecture using Coefficient Distributed Arithmetic. In: Proceedings of the IEEE Computer Society Annual Symposium on VLSI, May 11-12, 2005, pp. 162–166 (2005)

8. Aggoun, A., Jalloh, I.: Two-dimensional DCT/IDCT architecture. In: IEE Proceedings Computers and Digital Techniques, January 2-10, 2003, vol. 150(1), pp. 2–10. IEEE Press, Los Alamitos (2003)
9. Baoyu, Z.: A New Algorithm for the 2D Discrete Cosine Transform. In: Fourth International Conference on Signal Processing Proceedings, October 12-16, 1998, pp. 85–88. IEEE Press, Los Alamitos (1998)
10. Samadi, P., Wu, H., Ahmadi, M.: The 2-D Quantized DCT With Distributed Arithmetic. In: Canadian Conference on Electrical and Computer Engineering, May 2006, pp. 2049–2052. IEEE Press, Ottawa (2006)
11. Gonzalez, R., Woods, C.: Digital Image Processing, 3rd edn. Pearson, Prentice Hall (2007)

Wavelet-Packet Identification of Dynamic Systems with Coloured Measurement Noise

Henrique Mohallem Paiva¹ and Roberto Kawakami Harrop Galvão²

¹ Empresa Brasileira de Aeronáutica – EMBRAER
São José dos Campos, SP, 12227-901, Brazil
henrique.paiva@embraer.com.br

² Instituto Tecnológico de Aeronáutica – ITA
CTA, São José dos Campos, SP, 12228-900, Brazil
kawakami@ele.ita.br

Abstract. This paper analyses the effect of coloured noise on a recently proposed technique for linear system identification in frequency subbands using wavelet packets. For this purpose, a simulation study involving the longitudinal dynamics of a flexible aircraft model is presented. The results reveal that the wavelet-packet identification outcome is robust with respect to changes in the spectral noise features. In particular, the identified frequency response is effectively smoothed in regions with poor signal-to-noise ratio. Finally, the results are favourably compared, in terms of resonance peak identification, with those obtained by standard time-domain identification methods.

1 Introduction

Identification techniques may be a useful alternative to physical modelling of dynamic systems when the underlying phenomena are complex or when the direct measurement of model parameters is not practical. In many cases, linear identification can generate adequate models for engineering purposes, provided that the system is operated within a suitable working range [12].

In recent years, there has been growing interest in applications of the Wavelet Transform for linear system identification. Wavelets have been employed to characterize spectrally rich impulse responses with long subsequences of small coefficients [9], which arise in problems such as echo canceling in communication networks. Moreover, wavelets have been used to represent both slow and fast parameter variations in the identification of time-varying systems, such as communication channels to be equalized [17]. A wavelet model particularly suitable for adaptive identification of linear periodic time-varying systems was presented in [6]. Several applications can also be found in the context of mechanical systems, such as modal parameter identification [7,11] and vibration signal modelling for fault diagnosis purposes [3].

It is worth noting that most research efforts concerning the use of wavelet decompositions in linear system identification have focused on multiscale octave-band and single-level M-band decomposition [15]. Wavelet-packet algorithms,

which are popular in other signal processing areas such as audio/image coding [14] and pattern recognition [10], have not been common in this context. This gap was addressed by a recent paper [13], in which wavelet packet decomposition trees were employed to establish frequency bands for the identification of subband models. The tree structure was optimized by using a generalized cross-validation method in order to achieve a compromise between accuracy and parsimony of the overall model. In comparison with a standard ARX (auto-regressive with exogenous input) identification method, the proposed technique was found to be superior in terms of resonance peak identification and sensitivity to white measurement noise. However, investigations concerning coloured noise were not carried out.

Appropriate treatment of coloured noise, which is the most common form of noise in practical situations [15], may be essential to obtain suitable models by system identification, as discussed elsewhere [8][16][18]. The present work investigates the effect of coloured measurement noise on the performance of the wavelet-packet identification algorithm proposed in [13]. For this purpose, a case study involving the dynamics of a flexible aircraft model is presented. Input-output data are generated by numerical simulations with the inclusion of additive measurement noise within a particular frequency region. The effect of the spectral noise features on the identification outcome is investigated by varying the frequency region in which the measurements are degraded. The wavelet-packet identification results are compared with those obtained by standard ARX and ARMAX (auto-regressive moving-average with exogenous input) methods.

The remaining sections of this paper are organized as follows. Section 2 briefly describes the wavelet-packet identification technique under consideration. The case study is presented in section 3. Finally, concluding remarks are given in section 4.

2 The Wavelet-Packet Identification Technique

This section presents a brief review of the wavelet-packet identification technique proposed in [13]. For this purpose, some notation details will be initially introduced. Let H and G be the lowpass and highpass filters associated to the wavelet-packet decomposition tree, respectively, and H_r and G_r be the associated reconstruction filters [15]. In the present work, these filters are adopted such that the filter bank is orthonormal. $\downarrow 2$ and $\uparrow 2$ denote downsampling and upsampling operations, respectively. \mathbf{y} is the signal to be decomposed. $\mathbf{x}_{i,j}\{\mathbf{y}\}$ represents the wavelet coefficients of signal \mathbf{y} at node (i, j) of the wavelet tree, where $j \geq 0$ indicates the resolution level and i is a sequential index ranging from 0 to $2^j - 1$. Node (i, j) is the parent of nodes $(2i, j+1)$ and $(2i+1, j+1)$, which are the children of node (i, j) . All nodes without children are called leaf nodes. The depth of the tree is defined as the highest level in which nodes are present. A tree is said to be complete if all leaf nodes are in the same resolution level. Scalars and row vectors are represented by italic and boldface lowercase symbols, respectively.

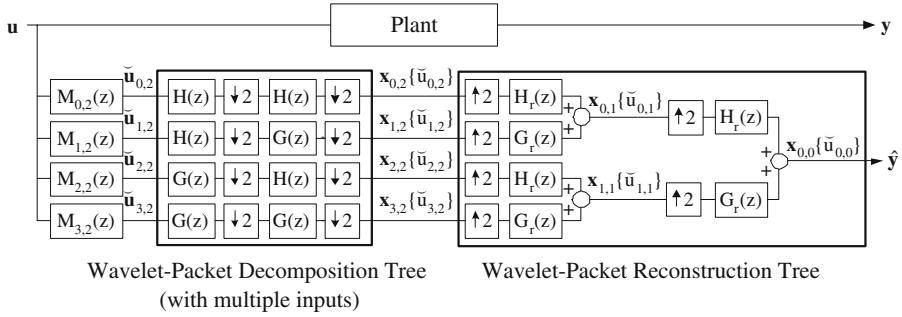


Fig. 1. Example of the system modelling scheme for a particular wavelet-packet decomposition tree with four leaf nodes (in this example, all leaf nodes are in the same level, but this is not a requirement)

The identification algorithm is based on the development of several subband models. The wavelet-packet decomposition tree is used to establish the frequency bands at which the subband models will be created. Each leaf node of the tree is associated to a frequency band, and the complete set of leaf nodes composes the whole frequency range. For each frequency band, a subband model is created.

This modelling scheme is illustrated in Fig. 1 for a particular structure of the wavelet-packet decomposition tree. In this figure, $M_{i,j}$ indicates the subband model intended to represent the plant in the frequency band associated to the leaf node (i, j) . If (i, j) is a leaf node, then signal $\check{u}_{i,j}$ is defined as the output of model $M_{i,j}$ for input u . If (i, j) is not a leaf node, then the coefficients $\mathbf{x}_{i,j}\{\check{u}_{i,j}\}$ are defined as the reconstruction of the coefficients at the children nodes of (i, j) .

The structure adopted for each subband model is a transfer function of the form of the form $M(z) = (1 - z^{-1})^{-s} (\alpha + \beta z^{-1})$, $s \in \mathbb{Z}$, $\alpha, \beta \in \mathbb{R}$.

Parameters $s_{i,j}$, $\alpha_{i,j}$ and $\beta_{i,j}$ are estimated in order to minimize the cost function $J_{i,j}(s_{i,j}, \alpha_{i,j}, \beta_{i,j}) = \mathbf{e}_{i,j}(\mathbf{e}_{i,j})^T$, where residue $\mathbf{e}_{i,j} = \mathbf{x}_{i,j}\{y - \check{u}_{i,j}\}$ denotes the wavelet-packet coefficients of the difference between the plant output y and the subband model output $\check{u}_{i,j}$, in the frequency band under consideration.

For a fixed value of $s_{i,j}$, cost $J_{i,j}$ is minimized with respect to $\alpha_{i,j}$ and $\beta_{i,j}$ by a least-squares procedure. A search algorithm is used to find the value of $s_{i,j}$ that leads to the minimum value of $J_{i,j}$. The details of these procedures are described in [13].

As demonstrated in [13], the cost $J_{0,0}$, associated to the root node $(i, j) = (0, 0)$ of the wavelet-packet tree, is equal to the square of the 2-norm of the prediction error $(y - \hat{y})$, that is, $J_{0,0} = (y - \hat{y})(y - \hat{y})^T$. Furthermore, for orthonormal wavelets, the cost at a non-leaf node (i, j) was shown to be equal to the sum of the costs at its children nodes, i.e., $J_{i,j} = J_{2i,j+1} + J_{2i+1,j+1}$. Such features allow the use of the following algorithm to choose the best wavelet-packet tree, which presents the minimum 2-norm of the prediction error [13].

1 - Fix the maximum depth d allowed for the tree and initialize the search with a complete tree with that depth.

2 - All nodes are candidates to be leaf nodes. Thus, for each node (i, j) , leaf or non-leaf, obtain the associated subband model $M_{i,j}(z)$. Calculate, for each node, the square of the 2-norm of the residue $\mathbf{e}_{i,j}$ and call this value $J_{i,j}^l$. Superscript l , standing for *leaf*, is used to emphasize that $J_{i,j}^l$ will be equal to the cost $J_{i,j}$ at node (i, j) only if this node is chosen to be a leaf.

3 - Initialize the costs at level d , the deepest level of the tree. If one node of this level is kept in the tree, it will be necessarily a leaf node. Thus, let $J_{i,d} = J_{i,d}^l$, for all $i = 0, 1, \dots, (2^d - 1)$.

4 - Analyze the other nodes of the tree, which can be made either leaf or non-leaf nodes. Start from level $j = d - 1$ and use a bottom-up approach (that is, analyze all nodes of a level before evaluating the parent level). Decide if each node (i, j) should be a leaf or a non-leaf node, by comparing cost $J_{i,j}^l$ (cost if (i, j) is a leaf node) with the sum $J_{2i,j+1} + J_{2i+1,j+1}$ of the costs at its children nodes (cost if (i, j) is a non-leaf node). The decision rule is:

$$J_{i,j} = \begin{cases} J_{i,j}^l, & \text{if } J_{i,j}^l \leq \rho(J_{2i,j+1} + J_{2i+1,j+1}) \\ J_{2i,j+1} + J_{2i+1,j+1}, & \text{otherwise} \end{cases} \quad (1)$$

where the penalty factor $\rho \geq 1$ ensures that node (i, j) will only be split into children nodes if the cost reduction is large enough to justify the increase in model complexity.

The penalty factor ρ is required to avoid an overfitting of the identification data. An increase in the value of ρ tends to reduce the number of nodes in the resulting tree. Thus, the choice of the penalty factor ρ can be regarded as a model order determination problem, which is addressed by using a generalized cross validation (GCV) method, described in detail in [13]. The value of ρ that minimizes the GCV index (GCVi) is selected, thus providing a tradeoff between model parsimony and identification accuracy. Therefore, the ability of the model to represent the behaviour of the system for input signals different from the one used in the identification is improved.

3 Case Study

This case study is concerned with the identification of a longitudinal flexible model for a simulated aircraft in straight flight [2]. This example was employed in [13] to illustrate the performance of the wavelet-packet technique in the presence of white measurement noise. The continuous-time transfer function adopted in the simulation is presented in equation (2), where the input and output variables indicate the elevator angle and the aircraft pitch angle measured by the sensor, respectively, both in degrees. In the simulation, this transfer function was discretized at a sampling rate of 40 Hz.

$$\frac{Y(s)}{U(s)} = \frac{0.4062s^3 - 2.034s^2 + 444.2s - 2015}{s^4 + 1.81s^3 + 1925s^2 + 1822s + 890400} \quad (2)$$

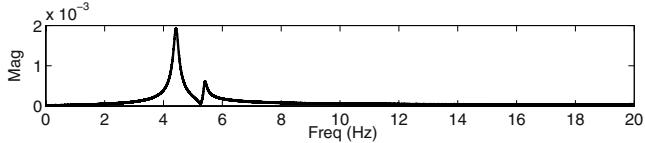


Fig. 2. Fast Fourier Transform of the plant output without measurement noise

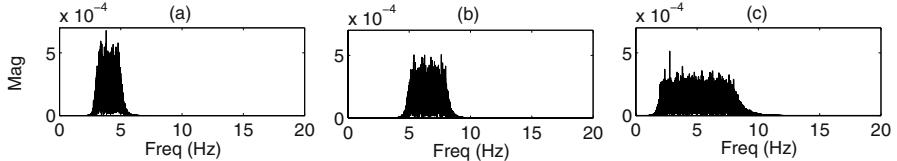


Fig. 3. Fast Fourier Transform of the measurement noise in the frequency bands (a) (3–5) Hz, (b) (5–8) Hz, and (c) (2–8) Hz

The following chirp excitation was used for identification: $u[kT] = \sin(2\pi f kT)$, where $k = 0, 1, \dots, 10^4$, $T = 0.025$ s, and frequency f varies linearly from 0.1 to 10 Hz. Fig. 2 presents the Fast Fourier Transform of the resulting output signal.

The *db8* filters from the Daubechies family were adopted, as in [13]. The maximum depth of the decomposition tree was set to six. For comparison, ARX and ARMAX model identifications were also carried out by using a standard prediction-error method [12]. The ARX model has general structure of the form: $y[k] = \sum_{i=1}^{n_a} a_i y[k-i] + \sum_{i=1}^{n_b} b_i u[k-i] + e[k]$, whereas the general structure of the ARMAX model is $y[k] = \sum_{i=1}^{n_a} a_i y[k-i] + \sum_{i=1}^{n_b} b_i u[k-i] + \sum_{i=0}^{n_c} c_i e[k-i]$, where e is a white noise process and $a_i, b_i, c_i \in \mathbb{R}$.

Parameters n_a and n_b were made equal to the true parameters of the system ($n_a = n_b = 4$). In this manner, the comparison with the wavelet-packet algorithm is carried out under favourable conditions for the ARX and ARMAX techniques. The sensor measurements were degraded with zero-mean white gaussian noise filtered by a 10th-order Butterworth bandpass filter. The ARMAX parameter n_c was set to 10 accordingly.

Three cases were considered, with measurement noise in the ranges (3–5) Hz, (5–8) Hz, and (2–8) Hz, respectively. The noise was normalized in order to have the same root-mean-square value in all three cases. Figs. 3a, 3b, 3c present the Fast Fourier Transform of a noise realization in each case. As can be seen, there is a significant overlap of the noise spectrum with respect to the system output shown in Fig. 2.

Fig. 4 depicts the wavelet-packet decomposition trees obtained according to the GCV criterion in the three noise cases. As can be seen, the tree with fewer nodes (Fig. 4c) corresponds to the situation in which the noise was spread over a wider frequency region (5 - 8 Hz). It can be argued that the GCV algorithm automatically joined the subbands with poor signal-to-noise ratio in order to increase the amount of data employed in the identification of each subband

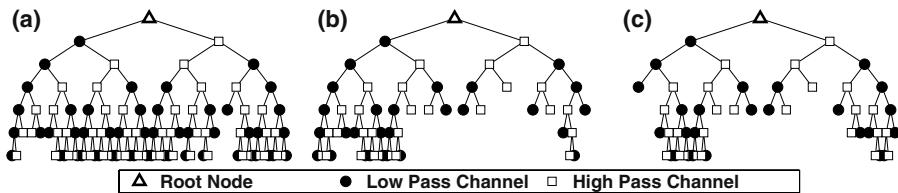


Fig. 4. Identification results: Wavelet-packet decomposition trees obtained for the three coloured-noise cases under study: (a) (3–5) Hz, (b) (5–8) Hz, (c) (2–8) Hz

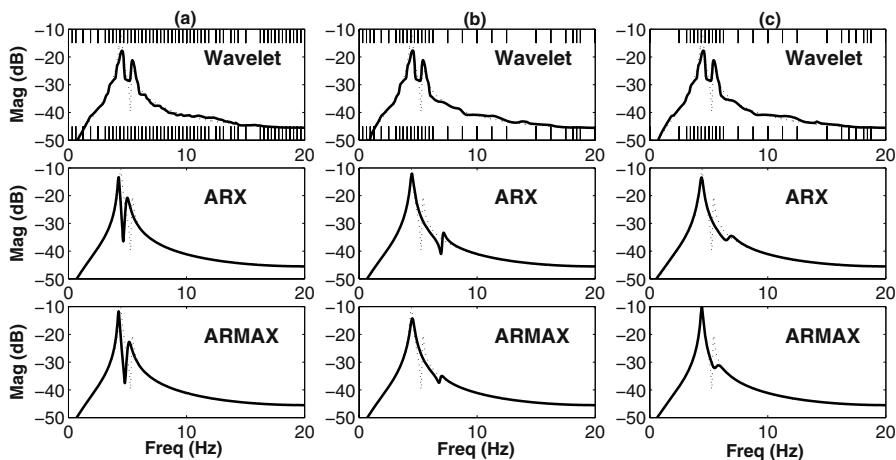


Fig. 5. Identification results: Wavelet-packet, ARX and ARMAX frequency responses obtained for the three coloured-noise cases under study: (a) (3–5) Hz, (b) (5–8) Hz, and (c) (2–8) Hz. The dotted curves indicate the exact frequency response of the system. In the wavelet-packet models, the vertical lines indicate the frequency partitions.

model. In contrast, the tree with more nodes (Fig. 4a) corresponds to the situation where the noise is concentrated on a frequency band (3 - 5 Hz) in which the plant output has the largest power, as shown in Fig. 2. In this case, joining subbands was not necessary because the signal-to-noise ratio was high throughout the entire spectrum.

Fig. 5 compares the frequency responses for the resulting wavelet-packet, ARX and ARMAX models. The graphs for the wavelet-packet models (top row) also indicate the subband partition of the frequency axis. As can be seen, wider frequency intervals are found in case (c), which is in agreement with the previous discussion concerning the decomposition trees. Fig. 5 also shows that the outcome of ARX identification is very sensitive to changes in the spectral features of the noise. Such a finding was expected because the ARX model structure assumes that the noise is white. The ARMAX results are slightly better, but the resulting model still fails to reproduce one of the resonance peaks in cases (b)

and (c). In contrast, the wavelet-packet algorithm correctly identified the position of both resonance peaks in all cases. Moreover, the algorithm can be seen to be robust with respect to changes in the noise spectrum, as the identification results in (a), (b), and (c) are almost indistinguishable.

It is worth noting that the values of n_a , n_b and n_c for the ARX and ARMAX models usually need to be estimated from the identification data. Therefore, in a more realistic identification scenario, the advantages of the wavelet-packet technique could be even more evident.

4 Conclusions

This work reviewed a recently proposed technique for linear system identification in frequency subbands by using wavelet packets. In the proposed formulation, the wavelet-packet decomposition tree is used to establish frequency bands where subband models will be created. An optimization of the tree structure is performed by using a generalized cross-validation method in order to achieve a compromise between accuracy and parsimony of the overall model.

In an application example involving the identification of the longitudinal flexible model of a simulated aircraft in straight flight, the wavelet-packet technique was favourably compared with standard ARX and ARMAX identification techniques in terms of resonance peak identification and sensitivity to coloured measurement noise. The present work shows that the wavelet-packet technique handles coloured noise in a natural manner by smoothing the identified frequency response in regions where the signal-to-noise ratio is poor.

Furthermore, it was observed that the wavelet-packet technique performs an adaptive partitioning of the frequency axis, which is an advantage of wavelet packets over the standard wavelet transform. It was shown that the proposed wavelet-packet technique leads to narrower frequency bands in the region of the frequency axis where the signal-to-noise ratio is larger.

Future works could adopt other basic structures for each subband model $M_{i,j}$. For instance, more filter taps could be used in the FIR term. Furthermore, the use of autoregressive terms could also be investigated. Finally, it would be worth studying the possibility of extending the technique to the identification of a broader class of systems, such as piecewise-affine plants.

Acknowledgements

The authors acknowledge the support of Embraer, CNPq (research fellowship and postdoctoral grant 200721/2006-2) and FAPESP (grant 2006/58850-6). The valuable suggestions of Prof. Takashi Yoneyama (Instituto Tecnológico de Aeronáutica) and Prof. Francisco Assis de Oliveira Nascimento (UnB - Universidade de Brasília) are also gratefully acknowledged.

References

1. Abdelghani, M., Verhaegen, M., Van Overschee, P., de Moor, B.: Comparison study of subspace identification methods applied to flexible structures. *Mechanical Systems and Signal Processing* 12(5), 679–692 (1998)
2. Burchales, A., Vacher, P.: Flexible aircraft model identification for control law design. *Aerospace Science and Technology* 6, 591–598 (2002)
3. Chen, H.X., Chua Patrick, S.K., Lim, G.H.: Adaptive wavelet transform for vibration signal modelling and application in fault diagnosis of water hydraulic motor. *Mechanical Systems and Signal Processing* 20(8), 2022–2045 (2006)
4. Coifman, R.R., Wickerhauser, M.V.: Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory* 32, 712–718 (1992)
5. Cooper, J.E., Desforges, M.J., Wright, J.R.: Modal parameter identification using an unknown coloured random input. *Mechanical Systems and Signal Processing* 9(6), 685–695 (1995)
6. Dorfan, Y., Feuer, A., Porat, B.: Modeling and identification of LPTV systems by wavelets. *Signal Processing* 84, 1285–1297 (2004)
7. Erlicher, S., Argoul, P.: Modal identification of linear non-proportionally damped systems by wavelet transform. *Mechanical Systems and Signal Processing* 21(3), 1386–1421 (2007)
8. Hasan, M.K., Chowdhury, A.K.M.Z.R., Khan, M.R.: Identification of Autoregressive Signals in Colored Noise Using Damped Sinusoidal Model. *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications* 50(7), 966–969 (2003)
9. Ho, K.C., Blunt, S.D.: Adaptive sparse system identification using wavelets. *IEEE Transactions on Circuits and Systems II - Analog and Digital Signal Processing* 49(10), 656–667 (2002)
10. Huang, K., Aviyente, S.: Information-theoretic wavelet packet subband selection for texture classification. *Signal Processing* 86(7), 1410–1420 (2006)
11. Huang, C.S., Su, W.C.: Identification of modal parameters of a time invariant linear system by continuous wavelet transformation. *Mechanical Systems and Signal Processing* 21(4), 1642–1664 (2007)
12. Ljung, L.: System Identification: Theory for the User, 2nd edn. Prentice-Hall, Englewood Cliffs (1999)
13. Paiva, H.M., Galvao, R.K.H.: Wavelet-packet identification of dynamic systems in frequency subbands. *Signal Processing* 86(8), 2001–2008 (2006)
14. Vera-Candeas, P., Ruiz-Reyes, N., Rosa-Zurera, M., Cuevas-Martnez, J.C., Lpez-Ferreras, F.: Sparse Approximations in Signal and Image Processing. *Signal Processing* 86(3), 432–443 (2006)
15. Vetterli, M., Kovacevic, J.: Wavelets and Subband Coding. Prentice-Hall, Upper Saddle River (1995)
16. Zheng, W.X.: Estimation of the Parameters of Autoregressive Signals From Colored Noise-Corrupted Measurements. *IEEE Signal Processing Letters* 7(7), 201–204 (2000)
17. Zheng, Y., Tay, D.B.H., Lin, Z.: Modeling general distributed nonstationary process and identifying time-varying autoregressive system by wavelets: theory and application. *Signal Processing* 81, 1823–1848 (2001)
18. Zheng, W.X.: Parametric Identification of Linear Systems Operating Under Feedback Control. *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications* 48(4), 451–458 (2001)

Comparison of the Cross Deleted Wigner Representation and the Matching Pursuit Distribution (Via Adaptive Signal Decomposition)

S. Ghofrani¹ and D.C. McLernnon²

¹ Electrical Engineering Department, Islamic Azad University, Tehran South Branch,
Tehran, Iran

² School of Electronic and Electrical Engineering, the University of Leeds, Leeds, UK
s_ghofrani@azad.ac.ir

Abstract. The cross-deleted Wigner representation (CDWR) suppresses the cross terms present in the Wigner distribution of a multi-component signal by decomposing the Wigner distribution (via the Gabor expansion) into terms that affect the auto- and cross-terms. The Matching Pursuit (MP) distribution is also a time-frequency representation that is devoid of cross terms. In this paper we decompose a signal by both the Gabor expansion and MP. Then the CDWR and the MP distribution are obtained. We then compare and contrast both representations (along with the original Wigner-Ville distribution (WVD)) with respect to (i) concentration/resolution, (ii) noise reduction capabilities and (iii) frequency and time resolvability.

Keywords: Wigner-Ville distribution; cross deleted Wigner representation; matching pursuit; Gabor expansion.

1 Introduction

Time-frequency distributions have been used extensively for non-stationary signal analysis. They describe how the frequency content of a signal changes with time. The Wigner-Ville distribution (WVD) is the best known of these [1]. For the signal $x(t)$ the WVD is:

$$WVD_x(t, \omega) = \int_{-\infty}^{+\infty} x\left(t + \frac{\tau}{2}\right) x^*\left(t - \frac{\tau}{2}\right) e^{-j\omega\tau} d\tau \quad (1)$$

where '*' represents complex conjugate and $\omega = 2\pi f$. The WVD satisfies an exceptionally large number of desirable mathematical properties, and in addition it has the best resolution among all other time-frequency distributions in the Cohen class. However, it generates cross terms when analyzing multi-component signals. These cross terms (or artefacts) mean that the distribution shows energy which does not really exist at particular time/frequency co-ordinates.

The signal decomposition is an approach to avoid the cross-terms from the time-frequency plane. In general the WVD of a decomposed signal includes two terms: the auto-Wigner and the cross-Wigner terms. The cross-deleted Wigner representation (CDWR) [2], [3] and the Matching Pursuit (MP) [4] are created by just considering

the auto-Wigner terms. Obviously the method of signal decomposition directly affects the extracted time-frequency representation.

Dennis Gabor first suggested expanding a signal onto a set of functions that are concentrated in both the time and frequency domains [5]. Later, Mallat and Zang also suggested a similar approach [4]. But the two decomposition methods have fundamental differences. In the Gabor transform we use constant band-limited functions and project a signal onto the regular lattice network. But in MP decomposition, except for the first iteration, we find the atom most similar to the residue. Also, in MP decomposition the time and frequency bandwidth of the elementary functions can be changed in order to get the most similarity with the signal or the residue.

In this paper we decompose the signal via both the non-adaptive Gabor approach and the adaptive MP implementation. We then compute the CDWR and the MP distribution according to the extracted expansion coefficients. The CDWR is compared with the MP distribution with respect resolution, noise reduction capability, and frequency/time resolvability.

So first in section 2 the Gabor transform is explained. In section 3 MP decomposition theory is reviewed. In section 4 we analyze a non stationary synthesized signal according to the two different decomposition methods and finally in section 5 we draw some conclusions.

2 Review of (Non-adaptive Signal Decomposition) Gabor Transform

For signal $x(t)$ the Gabor expansion is defined as [6]:

$$x(t) = \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} c_{m,n} h_{m,n}(t) \quad (2)$$

where $h_{m,n}(t) = h(t - mT)e^{jn\Omega t}$, $m, n \in \mathbb{Z}$, and T and Ω represent the time and frequency sampling intervals respectively. The synthesis function, $h(t)$, is subject to a unit energy constraint. The existence of (2) has been found for arbitrary $x(t)$ only for $T\Omega \leq 2\pi$. Generally, $h_{m,n}(t)$ do not form an orthogonal basis. This means that the first problem in using the Gabor expansion is how the coefficients should be determined. The first solution is to introduce an auxiliary function. The auxiliary function, $\gamma(t)$, and the synthesis function, $h(t)$, are named biorthogonal if:

$$\langle h_{m,n}, \gamma_{p,q} \rangle = \int_{-\infty}^{+\infty} h(t - nT) \gamma^*(t - qT) \exp(j2\pi(m-p)\frac{t}{T}) dt = \delta_{n-q} \delta_{m-p}. \quad (3)$$

If we can find $\gamma(t)$ then the Gabor coefficient $c_{m,n}$ can be computed as follows:

$$c_{m,n} = \langle x(t), \gamma_{m,n}(t) \rangle = \int_{-\infty}^{+\infty} x(t) \gamma_{m,n}^*(t) dt \quad (4)$$

where $\gamma(t)$ is also called the *analysis function*. The second problem in using the Gabor expansion is whether the coefficients reflect the signal behavior or not. For this

purpose the synthesis function and the analysis function must to be localized in both the time and frequency domains and have the most possible similarity to each other. Gabor chose the Gaussian pulse because it is a unique function which exhibits maximum concentration in both time and frequency domains.

Now the WVD of the decomposed signal in (2) is written as:

$$WVD_x(t, \omega) = \sum_{m,n=-\infty}^{+\infty} |c_{m,n}|^2 W_{h_{m,n}}(t, \omega) + \sum_{m,n=-\infty}^{+\infty} \sum_{m',n'=-\infty}^{+\infty} c_{m,n} c_{m',n'}^* W_{h_{m,n} h_{m',n'}}(t, \omega) \quad (5)$$

where $W_{h_{m,n}}(t, \omega)$ and $W_{h_{m,n} h_{m',n'}}(t, \omega)$ are the auto- and cross-WVD of the appropriate synthesis functions.

In this way, Qian, and Morris [2] introduced the time- frequency distribution series while Kadamb and Orr suggested [3] that the double sum in (5) corresponds to the cross-terms of the WVD. So keeping only the first term in (5), they defined the CDWR as follows:

$$CDWR_x(t, \omega) = \sum_{m,n=-\infty}^{+\infty} |c_{m,n}|^2 W_{h_{m,n}}(t, \omega). \quad (6)$$

3 Review of (Adaptive Signal Decomposition) Matching Pursuit

Mallat and Zhang [4] proposed an adaptive signal decomposition method. This method is based on a dictionary that contains a family of elementary functions or *time-frequency atoms*. The decomposition of a signal is performed by projecting the signal over the function dictionary and then selecting the atoms which can best match the local structure of the signal. So we compute a linear expansion of $x(t)$ over a set of elementary functions selected from the dictionary in order to best match its inner structures. The MP decomposition after M iterations can be written as follows:

$$x(t) = \sum_{n=0}^{M-1} c_n g_{\gamma_n}(t) + R^M x(t) \quad (7)$$

where $x(t)$ is the decomposed signal and $R^M x(t)$ is the residue after M times signal decomposition. By letting $R^0 x(t) = x(t)$, then the MP algorithm decomposes the residue at each stage. Thus the original signal is projected onto a sum of elementary functions which are chosen to best match its residues. Now $g_{\gamma_n}(t)$ is the *time- frequency atom* that belongs to the dictionary and which satisfies the unit-norm requirement – i.e.

$\int_{-\infty}^{+\infty} |g_{\gamma_n}(t)|^2 dt = 1$; the coefficient $c_n = \langle R^n x(t), g_{\gamma_n}(t) \rangle = \int_{-\infty}^{+\infty} R^n x(t) \cdot g_{\gamma_n}^*(t) dt$ is the inner

product of the functions $R^n x(t)$ and $g_{\gamma_n}(t)$; and γ_n refers to the atom's parameter set. So when the number of iterations is infinitive, then the residue will be zero, and we can say

$$\lim_{M \rightarrow \infty} R^M x(t) = 0 \Rightarrow x(t) = \sum_{n=0}^{+\infty} c_n g_{\gamma_n}(t) \quad (8)$$

where c_n is complex. The WVD of the decomposed signal is thus written as:

$$WVD_x(t, \omega) = \sum_{n=0}^{+\infty} |c_n|^2 W_{g_{\gamma_n}}(t, \omega) + \sum_{n=0}^{+\infty} \sum_{m=0, m \neq n}^{+\infty} c_n c_m^* W_{g_{\gamma_n} g_{\gamma_m}}(t, \omega) \quad (9)$$

where $W_{g_{\gamma_n}}(t, \omega)$ and $W_{g_{\gamma_n} g_{\gamma_m}}(t, \omega)$ are the auto- and cross-WVD of the appropriate atoms. Mallat and Zhang [4] suggested that the double sum in (9) corresponds to the cross terms of the WVD. So keeping only the first term in (9), they defined the MP distribution as follows:

$$E_x(t, \omega) = \sum_{n=0}^{+\infty} |c_n|^2 W_{g_{\gamma_n}}(t, \omega). \quad (10)$$

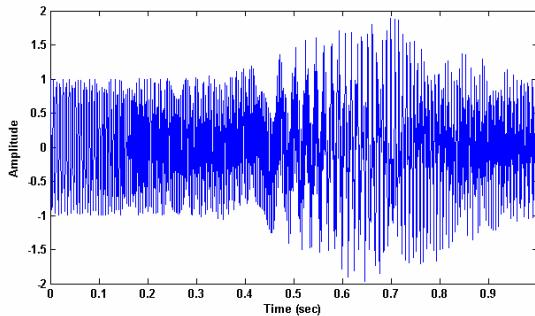


Fig. 1. The real part of the non-stationary signal in (11)

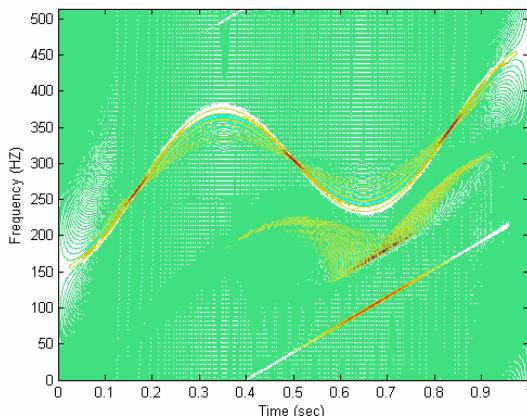


Fig. 2. The Discrete WVD of (11), with excellent resolution and negligible cross terms

This new distribution, $E_x(t, \omega)$, can now be interpreted as an energy density function of $x(t)$ in the time-frequency plane. Mallat and Zhang claimed that this new distribution does not include any cross terms.

4 Analyzing Synthetic Test Signals

Now we consider the following signal:

$$x(t) = e^{j(-60\sin(3\pi t) + 400t^2 + 1500t)} + e^{-25(t-0.67)^2} e^{j1200(t-0.4)^2} \quad (11)$$

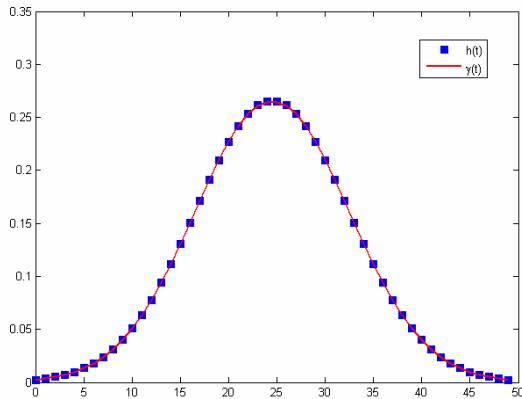


Fig. 3. The biorthogonal analysis and synthesis functions, both with same length $L=50$

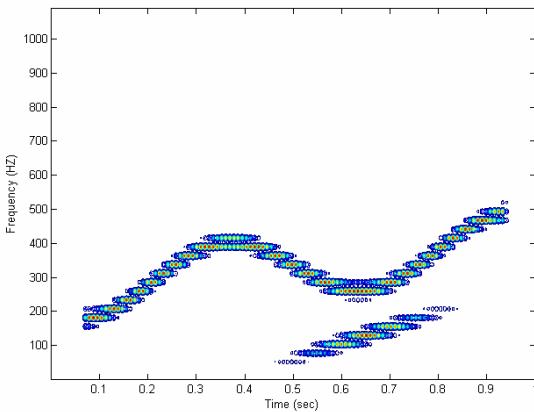


Fig. 4. The CDWR of $x(t)$ based on the Gabor expansion. The number of computed Gabor coefficients is 4210.

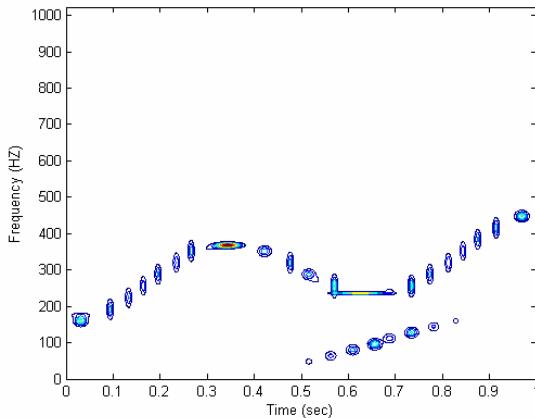


Fig. 5. The MP distribution of $x(t)$ based on Gaussian atoms. The number of iterations is 32.

This test signal (also previously used in [7]) is sampled at 1024 points over the time interval $[0, 1]$ and is plotted in Fig. 1. Its discrete WVD is then displayed in Fig. 2. We use an efficient algorithm [8] to obtain the discrete WVD where the window length has been chosen to give the best visual interpretation.

Now we decompose the signal via Gabor expansion theory and then construct the CDWR (see (6)). The major problem with the Gabor expansion is how to find the biorthogonal analysis function. In this paper we consider the orthogonal-like Discrete Gabor Transform (DGT) for an infinite signal [9] - as the real signal is usually very long sequence. In the orthogonal-like DGT, the Gabor coefficients can be thought of as the measure of similarity between the underlying signal, $x(i)$, and the individual basis functions, $\{h_{m,n}\}$. Thus it will reflect the signal's local behavior as long as the given synthesis window, $h(i)$, is indeed localized. So we try to obtain the biorthogonal function, $\gamma(i)$, that has the most similarity with $h(i)$. We restricted the length of $h(i)$ and $\gamma(i)$ to be the same. As long as the synthesis window function, $h(i)$, is localized, the orthogonal-like DGT will well reflect signal local behaviors because $c_{m,n}$ is very close to the inner product of $x(i)$ and $h_{m,n}$. Because the Gaussian window has the property of optimal concentration it is therefore very useful for time frequency analysis. We have chosen the synthesis window, $h(i)$, to be Gaussian with

length $L = 50$ and variance $\sigma^2 = \frac{N\Delta M}{2\pi}$, where ΔM denotes the discrete time sampling step and N is equal to the number of sampling points in frequency. The amplitude of the synthesis function is chosen as $A = \frac{1}{\sqrt[4]{\pi\sigma^2}}$ to satisfy the unit norm requirement. The optimum parameters for DGT have been chosen as:

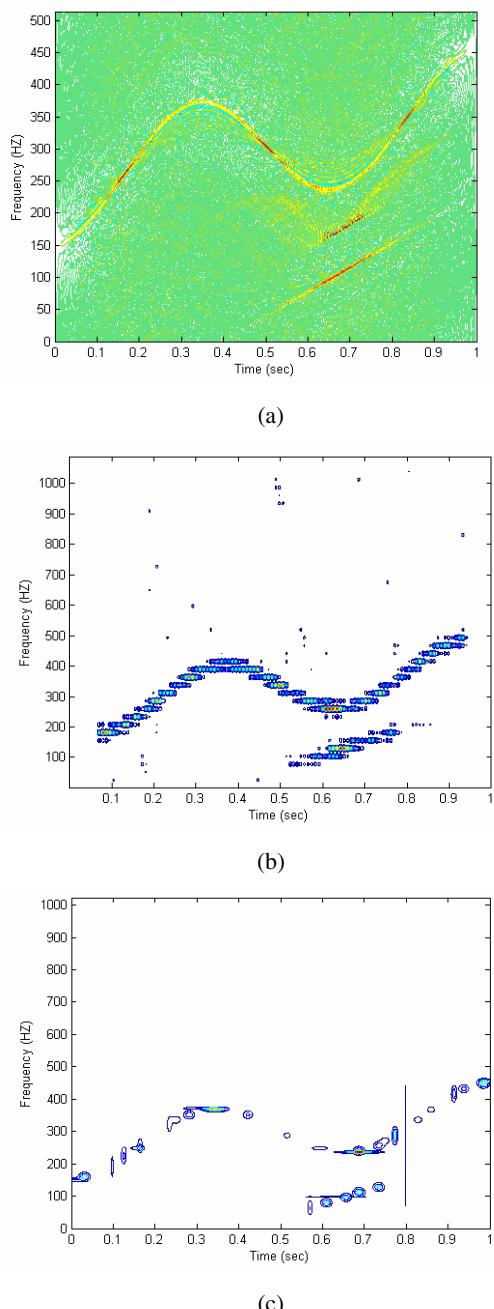


Fig. 6. (a) The Discrete WVD of the signal in (11) to which white Gaussian noise has been added ($\text{SNR}=0 \text{ dB}$); (b) the CDWR - number of Gabor coefficients is 4210; (c) the MP distribution - number of algorithm iterations is 32

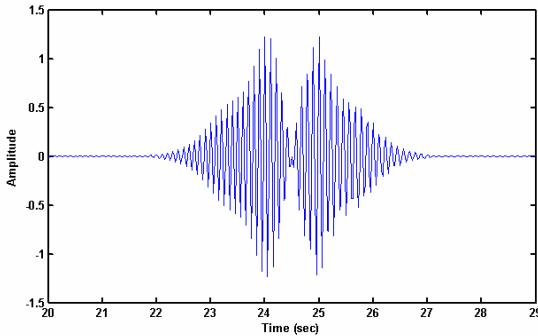


Fig. 7. The real part of the two component signal, in (14)

$$\Delta M = 10, N = 40, \Delta N = \frac{L + L_x}{N} . \text{ Note that } L_x = 1024 \text{ refers to the signal length, and}$$

ΔN denotes the discrete frequency sampling steps. Therefore in order to implement the DGT we require $M \times N = 107 \times 40$ coefficients, where M is equal to the number of sampling points in time and determined according to the signal lengths and ΔM . So the synthesis window is selected to be Gaussian and the biorthogonal analysis function that has the most similarity to the synthesis function is determined according to the algorithm described in [9]. Fig.3 shows the calculated analysis and synthesis functions. Now we try to decompose the signal and obtain the CDWR. According to the M and N values, we have to project the signal on the $M \times N = 4210$ Gaussian functions. The whole Gaussian functions have the same time and frequency bandwidth and just have been shifted in time and or in frequency. The CDWR is shown in Fig. 4. No cross terms are there while decreasing the resolution in comparison with the discrete WVD, Fig. 2, is obviously. So we have also decomposed the signal according MP by using the Gaussian atoms. It is well known that the Gaussian atom is unique in the sense that it has the greatest “concentration” in both the time and frequency domains. Such an atom is:

$$g_{\gamma_n}(t) = \frac{1}{\sqrt{s_n}} g\left(\frac{t - u_n}{s_n}\right) e^{j\zeta_n t} \quad (12)$$

where $g(t) = 2^{1/4} e^{-\pi t^2}$ and $\gamma_n = (s_n, u_n, \zeta_n)$ represents the atom's parameters set. The parameter s_n controls the envelope width of g_{γ_n} . The parameters u_n and ζ_n are respectively the temporal placement and the frequency variable. The parameter set are all real. In addition, s_n is also positive. As we use the Gaussian dictionary for the MP signal decomposition, the index n refers to the different atoms that exist in the dictionary. Now, suppose that a signal $x(t)$ is decomposed adaptively by employing the Gaussian elementary functions as in (8). It can be demonstrated that the MP distribution with the Gaussian atom is:

$$E_x(t, \omega) = 2 \sum_{n=0}^{+\infty} |c_n|^2 \tilde{H}\left(\frac{t - u_n}{s_n}\right) \tilde{F}(s_n(\omega - \zeta_n)) \quad (13)$$

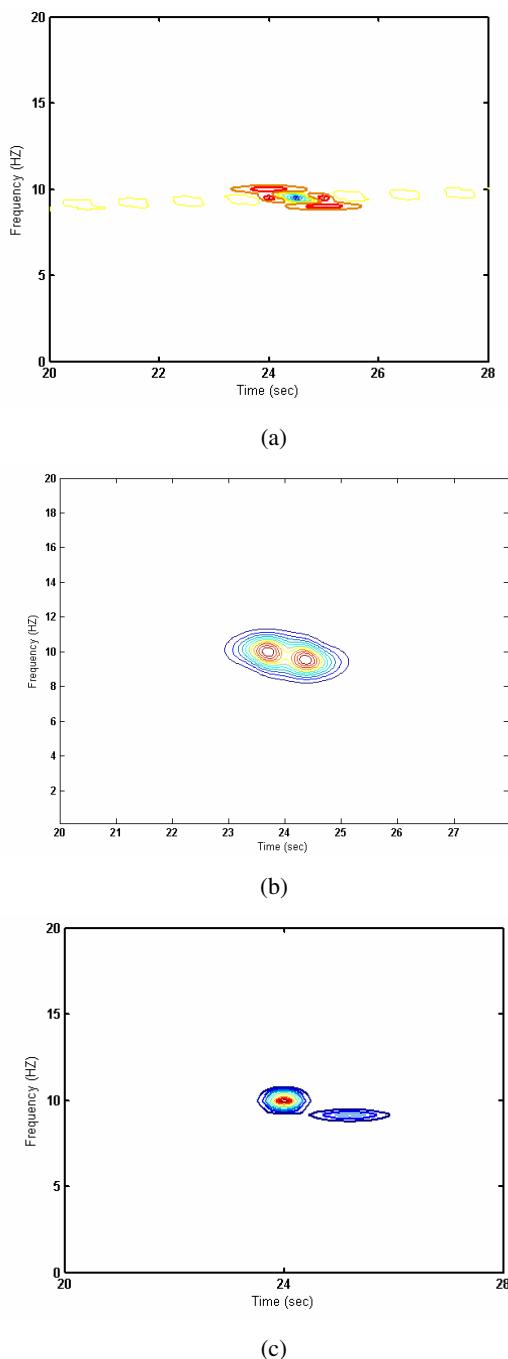


Fig. 8. (a) The Discrete WVD of the two-component signal in (14); (b) the CDWR; and (c) the MP distribution. It can be seen that the time and frequency resolvability of the MP is better than both the discrete WVD and the CDWR.

with $\tilde{H}(t) = e^{-2\pi t^2}$ $\tilde{F}(\omega) = e^{-\frac{1}{2\pi}\omega^2}$. Although we have explained the MP decomposition theory for Gaussian elementary functions for the continuous time signals, it should be implemented for processing discrete time signals and so we use the algorithm in [10]. Thus we have decomposed the signal $x(i)$ and obtained the best 32 Gaussian atoms. The MP distribution has been evaluated and is illustrated in Fig. 5. As we can see, there are no significant cross terms. The MP distribution has also better resolution compared to the CDWR. In addition the 4210 coefficients must be computed via Gabor decomposition while the number of MP algorithm iterations is set to be just 32. We also added white Gaussian noise (SNR=0 dB) to the signal in (11) and determined the CDWR according to the Gabor coefficients and the MP distribution via using the Gaussian elementary function, as well as the discrete WVD. These results are shown in Fig. 6. Note that the noise is spread in time-frequency plane for both the discrete WVD and the CDWR. Note that the noise affects the choice of the optimum atoms for the MP distribution. This feature may be interpreted as the MP noise reduction capability.

For testing resolvability, we consider the following $x(t)$:

$$x(t) = 4\sqrt{\frac{2}{\pi}} \cdot e^{-(t-24)^2} e^{j20\pi(t-24)} + 4\sqrt{\frac{2}{\pi}} \cdot e^{-(t-25)^2} e^{j18\pi(t-25)}. \quad (14)$$

The signal is sketched in Fig. 7. It includes of two Gaussian components. The first term is concentrated at (24, 10) in the time-frequency space and the second term is concentrated at (25, 9). The results are shown in Fig. 8. While the discrete WVD cannot distinguish the two terms at all, the CDWR shows one complex term with the two seeds. On the other hand the MP distribution detects and shows the two Gaussian components clearly. This feature is because of its capability to adapt both the atom's time and the frequency bandwidth in MP. So the adaptive decomposition shows a better performance with respect to the resolvability.

5 Conclusion

Although the potential of the WVD has long been recognized in signal processing, its applications are limited mainly due to the cross-terms interference problem. An alternative to the WVD is other time-frequency signal decompositions. One method is the constant band-width analysis achieved by the Gabor transform and the adaptive decomposition, MP. In this paper we have used the discrete WVD, CDWR, and the MP distribution for analyzing the non-stationary signals. The three time frequency representations are compared as regards resolution, noise reduction capability, and resolvability. Future work will concentrate on comparing the three time frequency representations according to both instantaneous frequency estimation and instantaneous bandwidth estimation.

References

1. Cohen, L.: Time-Frequency Distribution- A Review. Proceedings of the IEEE 77(7), 941–980 (1989)
2. Qian, S., Morris, J.: Wigner distribution decomposition and cross terms deleted representation. IEEE Trans. Sig. Proc. 27(2), 125–144 (1992)

3. Kadambé, S., Orr, R.: Instantaneous frequency estimation using the cross- term deleted Wigner representation (CDWR). In: TITS 1996, pp. 289–292 (1996)
4. Mallat, S.G., Zhang, Z.: Matching pursuit with time-frequency dictionaries. IEEE Trans. Signal Processing 41(12), 3397–3415 (1993)
5. Gabor, D.: Theory of communication. J. Inst. Electr. Eng. 93(3), 429–457 (1946)
6. Qian, S.: Introduction to time- frequency and wavelet transforms. Prentice-Hall, Englewood Cliffs (2002)
7. Qian, S., Stankovic, L.: Analysis of noise in time- frequency distributions. IEEE Signal Processing Letters 9(9) (2002)
8. Boashash, B., Black, P.J.: An efficient real-time implementation of the Wigner-Ville distribution. IEEE Trans. ASSP 35(11), 1611–1616 (1987)
9. Qian, S., Chen, D.: Discrete Gabor Transform. IEEE Trans. Sig. Proc. 41(7), 2429–2438 (1993)
10. Ferrando, S.E., Kolasa, L.A.: Algorithm 820: a flexible implementation of matching pursuit for Gabor functions on the interval. ACM Trans. Mathematics Software 28(3), 337–353 (2002)

Mean Weight Behavior of Coupled LMS Adaptive Systems Applied to Acoustic Feedback Cancellation in Hearing Aids

Yasmín Montenegro M.¹ and José C. M. Bermudez²

¹ University of Antofagasta - Chile

² Federal University of Santa Catarina - Brazil

ymontenegrom@yahoo.com.mx, j.bermudez@ieee.org*

Abstract. This paper presents a transient analysis of a hearing aids adaptive feedback canceller. The system employs an LMS adaptive estimator and an LMS adaptive predictor operating simultaneously. The nature of the practical problem makes the input to the adaptive estimator and the interference to its output statistically correlated. First, a modification is proposed to the original structure that accelerates convergence without compromising the cancellation level. The modified structure is then analyzed for slow adaptation and for an autoregressive input signal. Analytical models are derived for the mean behavior of the adaptive weights. Monte Carlo simulations verify the accuracy of the derived model.

Keywords: Adaptive systems, feedback, LMS, hearing aids.

1 Introduction

About 10% of the world's population suffers from hearing loss [1] that represents a serious limitation. The most common remedy for the hearing-impaired is a hearing aid [1]. However, user benefits are still limited due to the lack of effective algorithms that exploit all the digital signal processing potential. The occlusion effect and the acoustic feedback are among the main user complaints. Acoustic feedback is perceived as a distortion of the desired signal, limiting the maximum hearing aid's stable gain. Fig. 1 shows the basic structure of an adaptive feedback canceller. The acoustic feedback path is modelled by w^0 . The adaptive filter $w(n)$ produces an estimate $\hat{y}(n)$ of the feedback signal $y(n)$. This estimate is subtracted from the microphone signal. $g(n)$ provides amplification. Ideally, $\hat{y}(n)$ converges to $y(n)$ for n large. Then, only the input signal $x(n)$ is preserved at the hearing aids input $e(n)$ so that $u(n) = g(n)x(n)$.

Several strategies have been proposed to control acoustic feedback in hearing aids [2,3,4,5,6]. Continuously adaptive feedback cancellation is the most promising strategy because feedback depends strongly on the acoustical environment.

* This work has been partially supported CNPq under grants Nos. 307024/2006-7 and 470792/2006-0, by CAPES and by the University of Antofagasta-Chile.

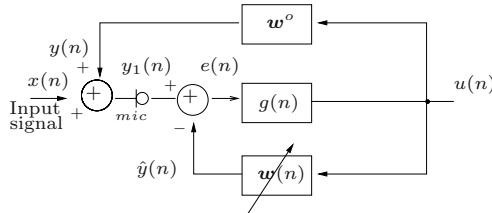


Fig. 1. Basic adaptive feedback cancellation in hearing aids

However, mean-square estimation of $y(n)$ in Fig. 1 leads to a biased solution because $x(n)$ and $u(n)$ are correlated. Several solutions proposed to reduce this bias introduce decorrelating operations such as delays or non-linearities in the forward path $g(n)$ or in the adaptive filter path [3][4]. The use of restrictions in the adaptive canceller has also been proposed [5][6]. Unfortunately, none of these solutions completely cancels the acoustic feedback.

Alternative solutions recently proposed rely on the direct method of closed-loop identification [7]. In this method, the data used for identification are obtained in closed loop but the identification is performed using an open loop model. Recent approaches assume a model for the input signal and apply the prediction-error method (PEM) to reduce the steady-state bias [8][9]. The performance of these systems has been studied only in steady-state. Though steady-state behavior is an important performance measure for adaptive systems, it does not provide sufficient information on which to base the system design. Transient algorithm behavior is crucial for real-time applications such as hearing aids. Most of the recent structures incorporate more than one adaptive filter operating simultaneously, and little knowledge is available about the transient behavior of such adaptive structures.

This paper studies the transient behavior of the adaptive hearing aid feedback cancellation system recently proposed in [9]. The adaptive filter (called shadow filter) works offline while a fixed cancellation filter operates in the actual signal path. The adaptive filter weights are periodically used to update the cancellation filter response. The adaptive filter updating structure includes an adaptive prediction error filter (PEF) to reduce the solution bias. Both adaptive filters are adapted simultaneously. A similar adaptive system has been analyzed in [10]. However, the analysis in [10] did not consider any additive interference to the unknown system output. In hearing aids, such interference is the desired signal $x(n)$, which is correlated to the adaptive filter input through the hearing aid amplifier. This feedback loop structure leads to a non-trivial transient analysis problem.

It is initially argued that the adaptive prediction error filter (PEF) should be applied to the output signal, instead of the estimation error as proposed in [9]. This change results in faster convergence without affecting steady-state performance. Then, an analytical model is derived for the mean behavior of the adaptive weights of the modified structure for autoregressive input signals and

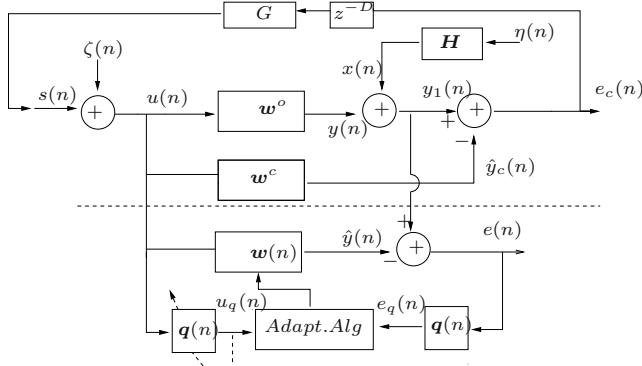


Fig. 2. Structure for feedback cancellation in hearing aids. Modified from [9]

under the assumption of slow adaptation. Monte Carlo simulation results show excellent agreement with the theoretical model.

2 Adaptive System Studied

Fig. 2 shows the adaptive feedback cancellation system studied. $x(n)$ is the desired signal. System \mathbf{H} and a white Gaussian noise $\eta(n)$ define a parametric model for $x(n)$. $\zeta(n)$ is a zero-mean white noise required for identifiability of the feedback path when $x(n)$ includes periodic components [9]. \mathbf{w}^o is the feedback path to be identified. $\mathbf{w}(n)$ is the adaptive estimation filter. $e_c(n)$ and $u(n)$ are, respectively, the input and output signals of the hearing aid, which is modeled by a gain G and a delay D . \mathbf{w}^c is the feedback estimation filter implemented in the signal path. \mathbf{w}^c is periodically updated with the coefficients of $\mathbf{w}(n)$ following some updating policy [9]. The adaptive prediction error filter $\mathbf{q}(n)$ whitens $u(n)$ before its use by the adaptive algorithm. The same filter $\mathbf{q}(n)$ is also applied to $e(n)$. Note that the top part (above broken line) of Fig. 2 operates as a stationary system during adaptation of the bottom part. Therefore, signals $u(n)$ and $y_1(n)$ are considered stationary during adaptation of $\mathbf{w}(n)$ and $\mathbf{q}(n)$.

The structure in Fig. 2 is the same proposed in [9] except for the adaptive predictor location. In [9] the adaptive predictor was applied to the error signal $e(n)$ and copied to filter $u(n)$. Here we propose the structure shown in Fig. 2. This solution pre-whitens the adaptive algorithm input signal and speeds-up convergence of stochastic gradient algorithms [10] without affecting steady-state behavior of the system.

Under certain conditions, the structure in [9] can provide an unbiased feedback path identification. If $x(n)$ is an autoregressive process, it is shown in [9] that $e(n)$ converges to $x(n)$ and the adaptive PEF (applied to $e(n)$) converges to \mathbf{H}^{-1} . From Fig. 2 and using the subscript q to refer to a signal filtered by $\mathbf{q}(n)$,

$$e_q(n) = y_{1q}(n) - \hat{y}_q(n). \quad (1)$$

Writing the PEF coefficient vector as $\mathbf{q}(n) = [1 \ \mathbf{p}^T(n)]^T$, $\hat{y}_q(n) = \hat{y}(n) + \sum_{i=1}^M p_i(n)\hat{y}(n-i) = \mathbf{u}_q^T(n)\mathbf{w}(n)$ and

$$e_q(n) = y_{1_q}(n) - \mathbf{u}_q^T(n)\mathbf{w}(n). \quad (2)$$

It is easy to show that $E[e_q^2(n)]$ is minimized for $\mathbf{w}(n) = \hat{\mathbf{w}}$ where $\hat{\mathbf{w}} = \mathbf{R}_{u_q u_q}^{-1} \mathbf{r}_q$ with $\mathbf{R}_{u_q u_q} = E\{\mathbf{u}_q(n)\mathbf{u}_q^T(n)\}$ and $\mathbf{r}_q^T = E\{y_{q1}(n)\mathbf{u}_q^T(n)\}$. Using $y_{q1}(n) = x_q(n) + \mathbf{u}_q^T(n)\mathbf{w}^o$ we have,

$$\hat{\mathbf{w}} = \mathbf{R}_{u_q u_q}^{-1} E\{x_q(n)\mathbf{u}_q(n)\} + \mathbf{w}^o \quad (3)$$

which shows that the solution bias is controlled by $E\{x_q(n)\mathbf{u}_q(n)\}$.

In Fig. 9 $\mathbf{q}(n)$ is adapted to whiten $e(n)$ and converges to the PEF for $x(n)$ upon convergence of $\mathbf{w}(n)$. Thus, $x_q(n)$ becomes uncorrelated to $u_q(n)$ in steady-state, yielding $E\{x_q(n)\mathbf{u}_q(n)\} = 0$ for $E\{x_q(n)\} = 0$, and the solution (3) becomes unbiased. It can be shown that the PEF applied to $u(n)$ as shown in Fig. 2 also converges to \mathbf{H}^{-1} . Thus, employing the adaptive PEF to either $e(n)$ or $u(n)$ leads to an unbiased steady-state solution. In steady-state both $e(n)$ and $u(n)$ are whitened by $\mathbf{q}(n)$. However, this is not true during transient because $\mathbf{q}(n)$ is the optimum PEF for $e(n)$ only upon convergence of $\mathbf{w}(n)$. On the other hand, the adaptive $\mathbf{q}(n)$ in Fig. 2 whitens $u(n)$ independently of $\mathbf{w}(n)$. Thus, the benefits of the PEF are available also during transient phase and improve overall convergence.

Fig. 3 shows Monte Carlo (MC) simulation results for the mean-square deviation (MSD) $E\{\|\mathbf{w}(n) - \mathbf{w}^o\|^2\}$ for both structures using the LMS algorithm for $\mathbf{w}^o = [-0.0016, 0.0016, 0.0046, 0.0502, -0.0691]^T$ and $H(z) = 1/(1 - 1.5z^{-1} + z^{-2} - 0.25z^{-3})$ [11]. Step sizes equal to 10^{-4} were used for both the LMS estimator and the LMS PEF. Also $\sigma_\eta^2 = 0.1875$ and $\sigma_\zeta^2 = 10^{-5}$. The faster convergence of the structure of Fig. 2 is clear.

3 Analysis

The following analysis considers the stationarity of the top part of the system in Fig. 2 and assumes the input signals stationary for the adaptation period.

3.1 Mean Weight Behavior for the Estimator

Consider orders N and M for the estimator $\mathbf{w}(n)$ and for the PEF $\mathbf{q}(n)$, respectively. Defining the $N \times 1$ weight error vector $\mathbf{v}(n) = \mathbf{w}(n) - \mathbf{w}^o$, the LMS update for the estimator in Fig. 2 is

$$\mathbf{v}(n+1) = \mathbf{v}(n) + \mu e_q(n)\mathbf{u}_q(n) \quad (4)$$

where $e_q(n) = \mathbf{q}^T(n)\mathbf{e}(n)$ with $\mathbf{e}(n) = [e(n), \dots, e(n-N+1)]^T$ and μ is the adaptation step.

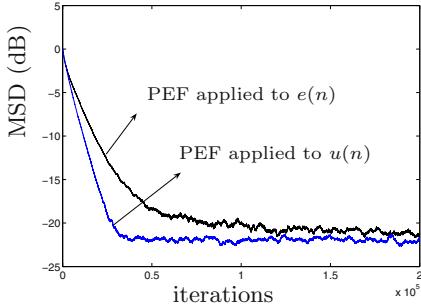


Fig. 3. Estimator MSD using the adaptive PEF on $e(n)$ and on $u(n)$

Assuming slow adaptation of both $\mathbf{q}(n)$ and $\mathbf{w}(n)$, we use the approximation $\mathbf{w}(n) \approx \dots \approx \mathbf{w}(n - N + 1)$ [10]. Under this approximation, $\mathbf{e}(n) = \mathbf{x}(n) - \mathbf{U}(n)\mathbf{v}(n)$ where $\mathbf{U}(n)$ is an $M \times N$ input vector matrix with columns $\mathbf{u}(n), \dots, \mathbf{u}(n - N + 1)$. Thus, $e_q(n)$ can be written as

$$e_q(n) = x_q(n) - \mathbf{u}_q^T(n)\mathbf{v}(n). \quad (5)$$

Using (5) in (4), taking the expectation and neglecting the correlation between $\mathbf{u}_q\mathbf{u}_q^T(n)$ and $\mathbf{v}(n)$ [12] yields

$$\begin{aligned} E\{\mathbf{v}(n+1)\} &= E\{\mathbf{v}(n)\} - \mu \overbrace{E\{\mathbf{u}_q(n)\mathbf{u}_q^T(n)\}}^{\alpha} E\{\mathbf{v}(n)\} \\ &\quad + \mu \underbrace{E\{x_q(n)\mathbf{u}_q(n)\}}_{\beta} \end{aligned} \quad (6)$$

where $\mathbf{u}_q(n) = \mathbf{u}(n) + \sum_{i=1}^M p_i(n)\mathbf{u}(n-i)$ and $x_q(n) = x(n) + \sum_{j=1}^M p_j(n)x(n-j)$ with $\mathbf{u}(n-i) = [u(n-i), \dots, u(n-i-M)]^T$. Neglecting the fluctuations of $\mathbf{p}(n)$ and the correlations between $\mathbf{p}(n)$ and both $u(n)$ and $x(n)$ (slow adaptation), the i th components of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, $i = 0, \dots, N-1$, in (6) are given by

$$\alpha_i(n) = \sum_{j=0}^{N-1} \sum_{k=0}^M \sum_{l=0}^M E\{q_k(n)\} E\{q_l(n)\} r_u(l+j-i-k) E\{v_j(n)\} \quad (7)$$

$$\beta_i(n) = \sum_{j=0}^M \sum_{k=0}^M E\{q_j(n)\} E\{q_k(n)\} r_{xu}(k+i-j) \quad (8)$$

where $r_u(l) = E[u(n)u(n-l)]$, $r_{xu}(l) = E[x(n)u(n-l)]$, $q_0(n) = 1$ and $q_\ell(n) = p_\ell(n)$ for $\ell = 1, \dots, M$. This completes the model for $E\{\mathbf{v}(n)\}$.

3.2 Mean Weight Behavior for the Predictor

The predictor LMS weight update equation is $\mathbf{p}(n+1) = \mathbf{p}(n) - \rho u_q(n) \mathbf{u}(n-1)$ with ρ as adaptation step.

Taking the expectation under the assumption of slow adaptation yields

$$\mathbb{E}[\mathbf{p}(n+1)] = [I - \rho \mathbf{R}_{uu}(0)] \mathbb{E}[\mathbf{p}(n)] - \rho \mathbf{r}_u(1) \quad (9)$$

for $\mathbf{R}_{uu}(0) = \mathbb{E}[\mathbf{u}(n-1)\mathbf{u}^T(n-1)]$ and $\mathbf{r}_u(1) = \mathbb{E}[u(n)\mathbf{u}(n-1)]$.

3.3 Correlations

Evaluation of $r_u(l)$ and $r_{xu}(l)$ is necessary to complete the analysis above. These correlations are obtained from the analysis of the stationary closed loop subsystem in the top part of Fig. 2 for an autoregressive input $x(n)$.

Autocorrelation $r_u(l)$. Defining the time-invariant weight error vector $\mathbf{v}^c = \mathbf{w}^c - \mathbf{w}^o$, we have

$$u(n) = \gamma(n) - G\mathbf{u}^T(n-D)\mathbf{v}^c \quad (10)$$

where $\gamma(n) = \zeta(n) + Gx(n-D)$. Taking the z-transform of (10) yields

$$H_u(z) = \frac{U(z)}{\Gamma(z)} = \frac{1}{1 + G\psi(z^{-1})\mathbf{v}^c z^{-D}} \quad (11)$$

with $\psi(z^{-1}) = [1, z^{-1}, z^{-2}, \dots, z^{-N+1}]^T$. Since $\zeta(n)$ is independent of $x(n)$,

$$R_u(z) = \frac{R_{\gamma u}(z)}{1 + G\psi(z^{-1})\mathbf{v}^c z^{-D}} \quad (12)$$

$$R_{\gamma u}(z) = \frac{R_\zeta(z)}{1 + G\psi(z)\mathbf{v}^c z^D} + \frac{G^2 R_x(z)}{1 + G\psi(z)\mathbf{v}^c z^D} \quad (13)$$

where $R_u(z)$ is the complex spectral density of $u(n)$ and $R_{\gamma u}(z)$ is the complex cross-spectral density of $u(n)$ and $\gamma(n)$.

For $x(n)$ autoregressive, $R_x(z) = H(z)H^*(1/z^*)\sigma_\eta^2$ with

$$H(z) = \frac{1}{\prod_{i=1}^M (1 + a_i z^{-1})}. \quad (14)$$

Assuming stability of (11) and using the Residue Theorem on (13), straightforward calculation leads to

$$\begin{aligned} r_{\gamma u}(l) &= \sigma_\zeta^2 \delta(l) + \sum_{k=1}^M \frac{G^2 \sigma_\eta^2 (-a_k)^M (-a_k)^{l-1}}{\prod_{\substack{i=1 \\ k \neq i}}^M (a_i - a_k) \prod_{j=1}^M (1 - a_j a_k)} \\ &\quad \times \frac{1}{(1 + G\psi(-a_k)\mathbf{v}^c(-a_k)^D)}, \quad l \geq 0. \end{aligned} \quad (15)$$

Finally, it follows from (12) that

$$r_u(l) = r_{\gamma u}(l) - G\mathbf{r}_u^T(l-D)\mathbf{v}^c \quad (16)$$

where $\mathbf{r}_u(l-D) = [r_u(l-D), \dots, r_u(l-D-N+1)]^T$.

Cross-correlation $r_{xu}(l)$. We have that $r_{xu}(l) = E\{x(n+l)u(n)\} = Gh(-l) * r_x(l+D)$ where $h(n)$ is the inverse z-transform of (14) and $*$ means convolution. Thus,

$$R_{xu}(z) = GH^*(1/z^*)R_x(z)z^D. \quad (17)$$

Substituting (14) and $R_x(z) = H(z)H^*(1/z^*)\sigma_\eta^2$ in (17) and using the Residue Theorem yields

$$\begin{aligned} r_{xu}(l) &= \sum_{k=1}^M \frac{G\sigma_\eta^2(-a_k)^{D+M}(-a_k)^{l-1}}{\prod_{\substack{i=1 \\ k \neq i}}^M (a_i - a_k) \prod_{j=1}^M (1 - a_j a_k)} \\ &\times \frac{1}{(1 + G\psi(-a_k)v^c(-a_k)^D)}, \quad l \geq 0 \end{aligned} \quad (18)$$

and

$$\begin{aligned} r_{xu}(l) &= \sum_{k=1}^M \frac{G\sigma_\eta^2(-a_k)^{D+M}}{(-a_k)^{l+1} \prod_{\substack{i=1 \\ k \neq i}}^M (a_i - a_k) \prod_{j=1}^M (1 - a_j a_k)} \\ &\times \frac{1}{(1 + G\psi(-a_k)v^c(-a_k)^D)}, \quad l < 0. \end{aligned} \quad (19)$$

4 Simulation Results

This section presents Monte Carlo (MC) simulations to verify the theoretical models. The parameters used in the example were $\mu = 0.001$, $\rho = 0.001$, $\sigma_\eta^2 = 1$ (using $\sigma_\eta^2 = 0.1875$), $\sigma_\zeta^2 = 10^{-5}$ and $H(z) = 1/(1 - 1.5z^{-1} + z^{-2} - 0.25z^{-3})$ (thus the optimal predictor is $p^o = [-1.5; 1; -0.25]^T$). Two feedback paths are utilized: $w_1^o = [-0.0016, 0.0016, 0.0046, 0.0502, -0.0691]^T$ and $w_2^o = [-0.0306, 0.0101, 0.0406, 0.0650, 0.0250]^T$ (first 5 samples of an actual feedback

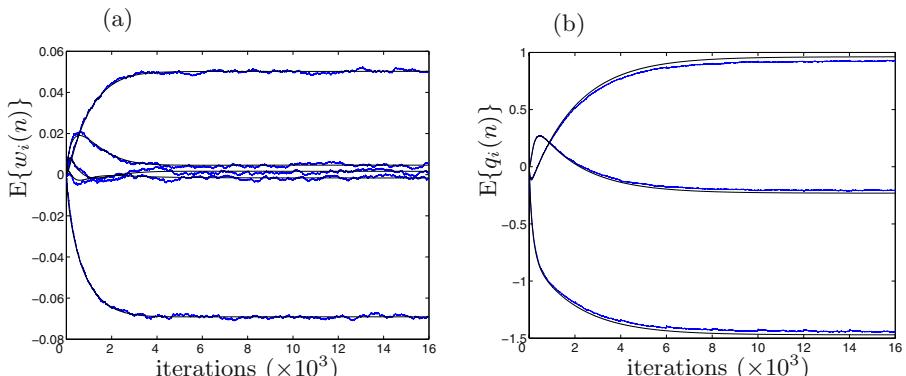


Fig. 4. Coefficient misadjustment in the filters: (a) Estimator and (b) Predictor for the feedback path w_1^o

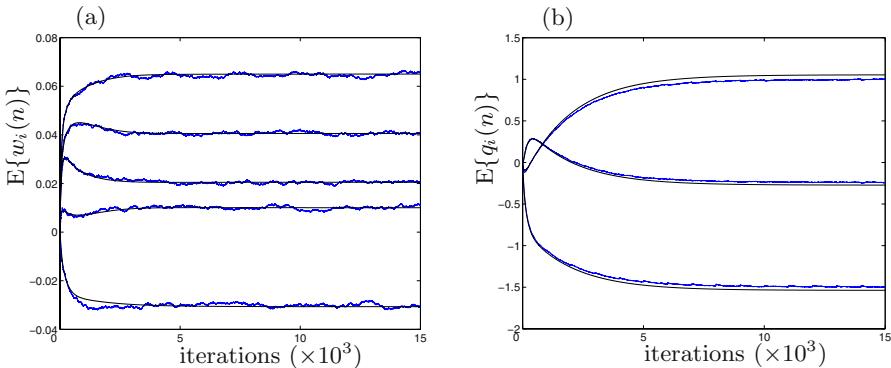


Fig. 5. Coefficient misadjustment in the filters: (a) Estimator and (b) Predictor for the feedback path w_2^o

path response). The Figs. 4 and 5 show the theoretical mean coefficient behaviors (smooth curves) and the MC simulations (ragged curves) averaged over 150 runs for the estimator and for the predictor. These results show excellent agreement between simulation results and the theoretical model provided by (6), (9), (16), (18) and (19) both in the transient and in the steady-state phases of adaptation.

5 Conclusions

This paper has presented a transient analysis of two coupled adaptive systems, one LMS estimator and one LMS predictor, when applied to feedback cancellation in hearing aids. A modification was proposed to the structure of [9]. The modification increases the convergence speed of the estimator filter without compromising the steady-state performance. Analytical models for first moment of both the adaptive filter weight vectors were derived. Monte Carlo simulation results illustrate the accuracy of the new models both during transient and in steady-state.

References

- Chung, K.: Challenges and recent developments in hearing aids part.1 speech understanding in noise, microphone technologies and noise reduction algorithms. *Trends In Amplification* 8(3), 83–124 (2004)
- Kates, J.M.: Feedback cancellation in hearing aids: results from a computer simulation. *IEEE Trans. on Speech and audio Processing* 39(3), 553–562 (1991)
- Joson, H.A.L., Asano, F., Toshio, Y.S.S.: Adaptive feedback cancellation with frequency compression for hearing aids. *J. Acoust. Society of America* 94(6), 3248–3254 (1993)
- Siqueira, M., Alwan, A.: Steady-state analysis of continuous adaptation in acoustic feedback reduction systems for hearing-aids. *IEEE Trans. on speech and audio processing* 8(4), 443–453 (2000)

5. Kates, J.M.: Constrained adaptation for feedback cancellation in hearing aids. *J. Acoust. Society of America* 106(2), 1010–1019 (1999)
6. Benesty, J., Huang, Y.: *Adaptive Signal Processing: Applications to Real-World Problems*. Springer, New York (2003)
7. Ljung, L.: *System identification, theory for the user*. Prentice Hall, Englewood Cliffs (1987)
8. Hellgren, J., Forssell, U.: Bias of feedback cancellation algorithms in hearing aids based on direct closed loop identification. *IEEE Trans. Speech and audio processing* 9(7), 906–913 (2001)
9. Spiet, A.I., Proudler, M., Wouters, J.: Adaptive feedback cancellation in hearing aids with linear prediction of the desired signal. *IEEE Trans. on Signal processing* 53(10), 3749–3763 (2005)
10. Mboup, M.B., Bershad, N.: Lms coupled adaptive prediction and system identification: a statistical model and transient mean analysis. *IEEE Transaccion Signal Processing* 42(10), 2607–2615 (1994)
11. Ho, K.C., Blunt, S.D.: Rapid identification of a sparse impulse response using an adaptive algorithm in the haar domain. *IEEE Trans. on Signal Processing* 51(3), 628–638 (2003)
12. Minkoff, J.: Comment: On the unnecessary assumption of statistical independence between reference signal and filter weights in feedforward adaptive systems. *IEEE Trans. on Signal Processing* 49(5), 1109 (2001)

Sound Localization Based on Excitation Source Information for Intelligent Home Service Robots

Keun-Chang Kwak

Dept. of Control, Instrumentation, and Robotic Engineering, Chosun University
375 Seosuk-dong, Dong-gu, Gwangju, 501-759, Korea
kwak@chosun.ac.kr

Abstract. This paper is concerned with Sound Localization (SL) using Excitation Source Information (ESI) and effective angle estimation for intelligent home service robots that are equipped with multi-channel sound board and three low-cost condenser microphones. The main goal is to localize a caller by estimating time-delay with features obtained from the excitation source based on Linear Prediction (LP) residual and Hilbert envelop, when the speaker calls robot's name in all directions. For performance analysis, we collected SL-DB (sound localization database) with the variation of distance and angle under test-bed environments like home. Here the localization success rate (LSR) and average localization error (ALE) from field of view (FOV) range of robot camera are used as localization performance criterion. The experimental results reveal that the presented method shows a good performance in comparison with the well-known Time Delay of Arrival (TDOA) and Generalized Cross Correlation-Phase Transform (GCC-PHAT) method.

Keywords: Sound localization, excitation source information, intelligent home service robots, effective angle estimation, low-cost microphones.

1 Introduction

The recent years have been witnessed a considerable number of studies on Sound Localization (SL) for Human-Robot Interaction (HRI) under intelligent robot environments. Based on this technique, the robot can move and help for giving aid to a person by recognizing and judging a situation in public places and home. The most representative techniques frequently used in conjunction with SL method are Time Delay of Arrival (TDOA) [5] and Generalized Cross Correlation-Phase Transform (GCC-PHAT) [4][7]. The TDOA are widely used due to accuracy and simple computation. However, this method usually represents a poor localization performance under noise and reverberation environments. On the other hand, the GCC-based function is made more robust by deemphasizing the frequency-dependent weighting. However, the disadvantage of the PHAT weighting is that it places equal emphasis on both low and high SNR regions [1]. Various methods have been suggested for localization of speaker by modeling the production of speech. Most of the speech model-based methods use spectral features which correspond to the characteristics of the vocal tract system during the production of speech. The spectral features are affected by transmission through noise and room reverberation.

Recently, a few attempts have been performed with the aid of the characteristics of the excitation source during the production of speech in conjunction with speaker recognition [2] and speaker localization [1]. In this study, we present and discuss on SL method using Excitation Source Information (ESI) and effective angle estimation for intelligent home service robots equipped with multi-channel sound board called MIM(Multimodal Interface Module) in ETRI(Electronics and Telecommunications Research Institute) and low-cost condenser microphones. In contrast to typical SL method in robot environments, the underlying principle exploited here is to consider a three-phase development of robust SL method. First, we use Endpoint Detection (EPD) algorithm based on log energy on SL-DB (sound localization database) obtained from three microphones equipped in intelligent mobile robots [6]. Here the database is collected by the variation of distance and angle under test-bed environments like home. Next, we compute a time-delay between two speech signals using ESI on SL-DB. Finally, we estimate a reliable localization angle from several candidate angles obtained by time-delay estimation of ESI. This paper is organized as follows. Section 2 describes the previous works on the well-known TDOA and GCC-PHAT frequently used in conjunction with SL method. In Section 3, we present a method for estimating time-delay using ESI. In Section 4, the SL-DB and the SL method based on ESI and effective angle estimation are described. Based on SL-DB, we perform comprehensive experiments in Section 5. Finally, conclusions are given in Section 6.

2 Previous Works: TDOA and GCC-PHAT

In this section, we describe the well-known previous works frequently used in conjunction with SL method such as TDOA and GCC-PHAT. These methods are compared with the presented technique on the constructed SL-DB in Section 5. Firstly we explain a brief review on time-delay estimation of TDOA. Let's consider windowed frames of N samples with 50% overlap. We assume that the index corresponding to each frame is omitted from the equations. It is necessary to define a coherence measure to determine time delay with the signals captured by two different microphones. The most well-known coherence measure is a simple cross-correlation between the signals perceived by two microphones as the following expression

$$R_{ij}(\tau) = \sum_{n=0}^{N-1} x_i[n]x_j[n-\tau] \quad (1)$$

where $x_i[n]$ is the signal received by i'th microphone and τ is the correlation lag in samples. The cross-correlation is maximal when τ is equal to the offset between the two received signals. However, this method is difficult to obtain a good estimate of the time-delay even when the signals are corrupted by noise and environments. On the other hand, GCC (Generalized Cross Correlation) is a correlation method in frequency domain. SL method based on GCC-PHAT has a lot of merits in noise environments and reverberation environments. When the signals $x_1(n)$ and $x_2(n)$ are

obtained by each of two microphones, the generalized cross-correlation between $x_1(n)$ and $x_2(n)$ can be obtained by the following equation.

$$R_{x_1x_2}(n) = \frac{1}{2\pi} \int_{-\infty}^{\infty} W(\omega) X_1(\omega) X_2^*(\omega) e^{j\omega n} d\omega \quad (2)$$

where $W(\omega)$ is a frequency weighting function. Among GCC-based methods, the most commonly used weighting function is PHAT (phase transform). Its weighting function is the reciprocal of $X_1(\omega)X_2^*(\omega)$. PHAT is a weighting function that determines the relative importance of each frequency as follows

$$W(\omega) = \frac{1}{|X_1(\omega)X_2^*(\omega)|} \quad (3)$$

The delay time between $x_1(n)$ and $x_2(n)$ can be obtained by the following expression

$$\tau = \arg \max R_{x_1x_2}(n) \quad (4)$$

3 Time-Delay Estimation of Excitation Source Information

In Linear Predication (LP) analysis, the sample $s(n)$ is estimated as a linear weighted sum of the past p samples. The predicted sample $\hat{s}(n)$ is given by

$$\hat{s}(n) = - \sum_{k=1}^p a_k s(n-k) \quad (5)$$

where p is the LP's order and $a_k, k = 1, 2, \dots, p$, is LP's coefficients. These coefficients are obtained by minimizing the mean squared error (MSE) between the predicted sample value and the actual sample value over the analysis frame (40ms). The error between the actual value $s(n)$ and the predicted value $\hat{s}(n)$ is computed as follows

$$r(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^p a_k s(n-k) \quad (6)$$

This error $r(n)$ is called the LP residual of the speech signal. The LP residual contains information about the excitation source. The values of LP residuals are large around the instants of glottal closure for voiced speech. It is difficult to derive information from short segments of LP residual due to large fluctuations in amplitude. Here the analytic signal $r_a(n)$ corresponding to $r(n)$ is given by

$$r_a(n) = r(n) + j r_h(n) \quad (7)$$

The strength of the LP residual at each instant is obtained by computing the Hilbert envelope of the LP residual signal. The Hilbert envelope of the residual signal is obtained by the following expression [1][3]

$$h_e(n) = |r_a(n)| = \sqrt{r^2(n) + r_h^2(n)} \quad (8)$$

Fig. 1 shows the results of EPD based on log energy. Here the speech signal of robot name used in this study is never. Fig. 2 visualizes the results obtained by LP residual and Hilbert envelope from two signals obtained through two microphones. Here we used only 250 ms (4000 samples) among the detected signal to perform fast computation. Fig. 3 shows the estimation results of the time-delay using ESI.

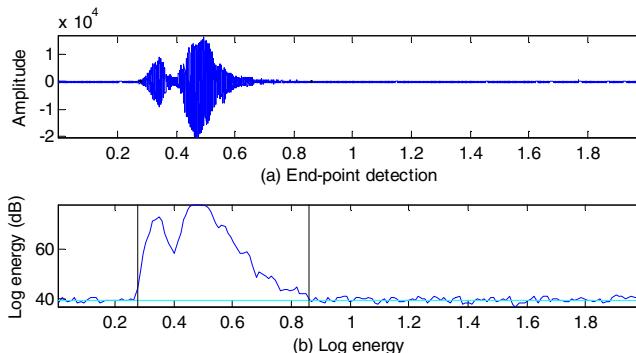


Fig. 1. Speech signal detected by EPD

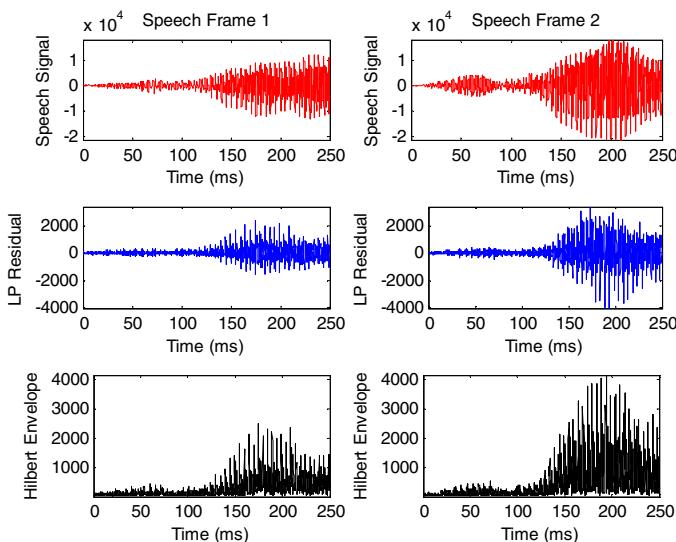


Fig. 2. LP residual and Hilbert envelope

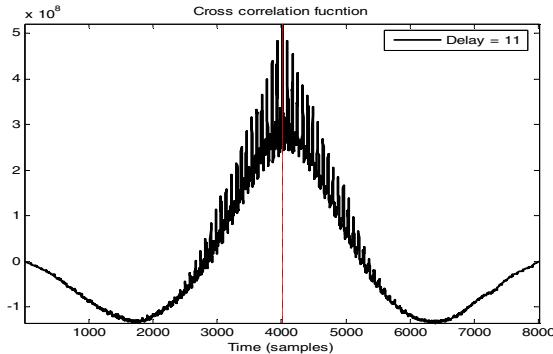


Fig. 3. Time-delay estimation by ESI

After computing time-delay, we obtain the localization angle between two microphones. Suppose that the sound wave at the microphone is a plan wave and the angle between microphones is 120° , respectively. Finally azimuth θ is estimated as follows

$$\theta = \cos^{-1}\left(\frac{\Delta t v}{L}\right) - 30 \quad (9)$$

where θ is the angle of sound source obtained in the above equation. Δt is time delay between two microphones and v is the velocity of sound source. Moreover L is the distance between two microphones.

4 Sound Localization Under Robot Environments

In this section, we describe SL-DB collected in test-bed environments like home. Furthermore, in order to localize sound source under robot environments, we present the method to estimate the effective localization angle from three time-delay values obtained by ESI.

4.1 SL-DB and Robot Environments

The SL-DB used in this study was constructed in test-bed environment that is similar with home environment to evaluate the SL algorithm. We used robot's name (wever) to collect SL-DB. The two individuals speak three times at 45° interval from 0° to 360° . The data set consists of 72 speeches at each meter from 1meter to 3 meter (M1 and M2 DB) when the number of low-cost condenser microphone is three. The audio is stored as a mono, 16bit, and 16kHz. Firstly we obtain the segmented speech signals from the database based on EPD algorithm. Fig. 4 shows the arrangement of three microphones with 120 degrees interval on Wever robot. The three microphones are equipped with multi-channel sound source board that is developed by ETRI. These microphones consist of low-cost condenser about four dollars, while the microphones presented in the previous literature are high-priced microphones.

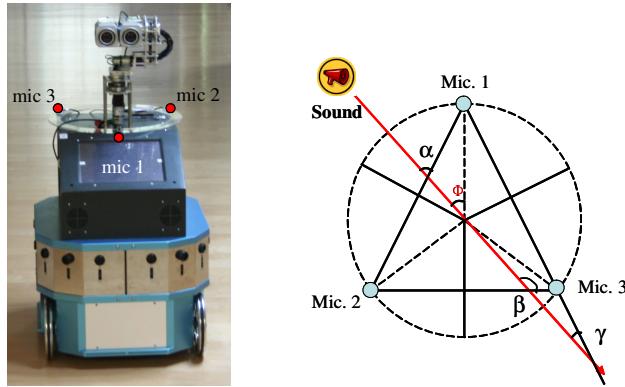


Fig. 4. Wever robot and the arrangement of microphones

4.2 Effective Angle Estimation

In what follows, we propose the method to estimate reliable localization angle from several candidate angles obtained by ESI. The three angles are obtained based on each time-delay obtained by ESI as the following equations

$$\alpha = \cos^{-1} \left(\frac{v \tau_{12}}{d} \right) \quad (10)$$

$$\beta = \cos^{-1} \left(\frac{v \tau_{23}}{d} \right) \quad (11)$$

$$\gamma = \cos^{-1} \left(\frac{v \tau_{13}}{d} \right) \quad (12)$$

where d (32cm) is a distance between each microphone and v (347.1 m/sec²) is the velocity of sound, τ_{12} is time delay between channel 1 and channel 2, τ_{23} is time delay between channel 2 and channel 3, τ_{13} is time delay between channel 1 and channel 3. From these angles, we obtain six candidate angles as follows

$$\begin{aligned} \Phi_1 &= \alpha - 30^\circ & \Phi_2 &= -\alpha - 30^\circ & \Phi_3 &= \beta + 90^\circ \\ \Phi_4 &= -\beta + 90^\circ & \Phi_5 &= \gamma + 30^\circ & \Phi_6 &= -\gamma + 30^\circ \end{aligned} \quad (13)$$

Here because it is difficult to obtain ideal time-delay, we have to select the two closest Φ . And then we estimate the reliable localization angle by averaging these two angles.

5 Experimental Results

In this section, the presented approach is compared with TDOA and GCC-PHAT on SL-DB. The localization success rate (LSR) and average localization error (ALE) are

considered as performance measure. The LSR is computed by FOV(± 15) of robot camera because SL method is used with face detection when robot moves toward caller. The experimental results are listed in Table 1. As summarized in table 1, the results of both LSR and ALE for M1 and M2 DB revealed that the presented method showed a better localization performance (more than 20%) in comparison to that of TDOA and GCC-PHAT. Here we only used 250 ms (4000 samples) period among whole speech signal detected by EPD algorithm. Table 2 lists six candidate angles and final estimated angle of the case of 0 degree in M1 data set.

Table 1. Performance comparison for M1 and M2 DB (during 250 ms, FOV ± 15)

	M1 set		M2 set	
	LSR (%)	ALE (degree)	LSR (%)	ALE (degree)
TDOA	79.2	5.44	63.9	5.29
GCC	56.9	5.97	81.9	4.53
The presented method	97.2	4.43	87.5	4.66

Table 2. Six candidate angles and final estimated angle (0 degree, 1~3 m)

	Φ_1	Φ_2	Φ_3	Φ_4	Φ_5	Φ_6	final angle
1m	11.77	-71.77	176.11	3.88	65.55	-5.55	7.83
	11.77	-71.77	176.11	3.88	65.55	-5.55	7.83
	11.77	-71.77	-176.11	-3.88	77.31	-17.31	-10.60
2m	11.77	-71.77	176.11	3.88	77.31	-17.31	7.83
	-1.80	-58.19	180.00	0	58.19	1.80	-0.90
	5.55	-65.55	180.00	0	65.55	-5.55	2.77
3m	-11.64	-48.35	-172.20	-7.79	48.35	11.64	-9.71
	11.77	-71.77	-176.11	-3.88	77.31	-17.31	-10.60
	-11.64	-48.35	180.00	0	58.19	1.80	0.90

6 Conclusions

We have developed SL method with the aid of ESI and effective angle estimation for intelligent home service robots. The experimental results regarding SL-DB used in this study revealed that the presented method showed a better localization performance in comparison to the TDOA and GCC-PHAT. As a result, we can naturally communicate face to face with home service robot through spontaneous speech

recognition with continuous words to provide useful information such as weather information and daily life schedule at a short distance. Furthermore, when somebody calls the robot's name at a long distance, home service robot can detect the direction of sound source in all directions and then move forward caller.

References

1. Raykar, V.C., Yegnanarayana, B., Prasanna, S.R.M., Duraiswami, R.: Speaker localization using excitation source information in speech. *IEEE Trans. on Speech and Audio Processing* 13(5), 751–761 (2005)
2. Murty, K.S.R., Yegnanarayana, B.: Combining evidence from residual phase and MFCC features for speaker recognition. *IEEE Signal Processing Letters* 13(1), 52–55 (2006)
3. Rao, K.S., Prasanna, S.R.M., Yegnanarayana, B.: Determination of instants of significant excitation in speech using Hilbert envelope and group delay function 14(10), 762-765 (2007)
4. Knapp, C.H., Carter, G.C.: The generalized correlation method for estimation of time delay. *IEEE Trans. on Acoustic, Speech, and Signal Processing ASSP-24*, 320–327 (1976)
5. Huang, J., Supaongprapa, T., Terakura, I., Wang, F., Ohnishi, N., Sugie, N.: A model based sound localization system and its application to robot navigation. *Robotics and Autonomous Systems*, 199-209 (1999)
6. Kwak, K.C., Kim, H.J., Bae, K.S., Yoon, H.S.: Speaker identification and verification for intelligent service robots. In: Int. Conference on Artificial Intelligence (ICAI 2007), Las vegas, pp. 515–519 (2007)
7. Park, B.C., Ban, K.D., Kwak, K.C., Yoon, H.S.: Sound source localization based on audio-visual information for intelligent service robots. In: The 8th Int. Symposium on Advanced Intelligent Systems (ISIS, Sokcho 2007), pp. 364–367 (2007)

Voice Detection in Noisy Speech Signal

Aïcha Bouzid¹ and Noureddine Ellouze²

¹ Institut Supérieur d'Electronique et de Communication de Sfax, Route Menzel Chaker
Km 0.5, B. P. 868, 3018 Sfax, Tunisia

² Ecole Nationale d'Ingénieurs de Tunis, BP. 37, Le Belvédère 1002 Tunis, Tunisia
bouzidach@yahoo.fr, N.Ellouze@enit.rnu.tn

Abstract. An algorithm for voicing detection in noisy speech signal is proposed. This algorithm is based on the product of wavelet transforms at some scales called multi-scale product. The multi-scale product has the ability to reinforce the edge in the signal while suppressing additive noise. Motivated by the fact that unvoiced sounds are, in most important speech production models, considered as filtered noise, we apply the multi-scale product on speech signal for detection of voiced segments. In fact, the multi-scale product anneals the signal frames corresponding to unvoiced sounds and frames of silence, while it conserves speech periodicity for voiced frames.

Keywords: Wavelet transform, multi-scale product, voicing decision, speech signal.

1 Introduction

Pre-processing of speech signal is very crucial in the applications where silence or background noise is completely undesirable. Applications like speech and speaker recognition [1] needs efficient feature extraction techniques from speech signal where most of the voiced part contains speech or speaker specific attributes. Silence removal is a well known technique adopted for many years for this and also for dimensionality reduction in speech that facilitates the system to be computationally more efficient. This type of classification of speech into voiced or silence/unvoiced sounds [2] finds other applications mainly in fundamental frequency estimation, formant extraction or syllable marking and so on.

There are several ways of classifying events in speech. It is accepted convention to use a three state representation in which states are (i) silence where no speech is produced; (ii) unvoiced sound, in which the vocal cords [3] are not vibrating, so the resulting speech waveform is random in nature and (iii) voiced sound, in which the vocal cords are tensed and therefore vibrate periodically when air flows from the lungs, so the resulting waveform is quasi-periodic [4]. It should be clear that the segmentation of the waveform into well defined regions of silence, unvoiced, signals is not exact; it is often difficult to distinguish a weak, unvoiced sound (like /f/ or /th/) from silence, or weak voiced sound (like /v/ or /m/) from unvoiced sounds or even silence.

The speech classification task has been studied in many articles by a variety of approaches since 1980's. Basically the classification is done by relying on different

types of feature vectors which are extracted from the input speech frames. These features can be derived by three approaches:

The first approach works in the time domain and uses statistical measurements. The common features are zero crossing rate, relative energy level, autocorrelation coefficients, etc. [5], [6], [7]. This approach only achieves good accuracy if using a large number of parameters.

The second approach works in the frequency domain. Frequently used features are the spectrum [8], optimal filters [9], Mel Frequency Cepstral Coefficients (MFCC) [10]. The results show that MFCCs give high performance.

The third approach combines both time and frequency domains, using the Short Time Fourier Transform (STFT) [11] or Wavelet Transform (WT) [12].

In this paper, we propose a time-scale approach using multi-scale product of wavelet transform for the detection of voicing frames in speech. The paper is organized as follows. The second section puts forward the time-scale analysis of signals. The third section concerns the edge detection of noisy signal by the multiplication of wavelet transform coefficients. Section 4 presents the proposed method for voicing detection. Section 5 concludes this work.

2 Wavelet Transform Modulus Maxima

Wavelet Transform is a linear powerful time-scale representation of signals, a real wavelet $\psi(t)$ which has the property of zero mean is used for this representation.

$$\int \psi(t) dt = 0. \quad (1)$$

The wavelet is delayed by a time factor u and scaled by a scale factor s , the basic expression of wavelet transform of the signal $f(t)$ is

$$Wf(u, s) = \int f(t) \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right) dt. \quad (2)$$

This expression of correlation can be written as a convolution product with the time reverse function $\bar{\psi}_s(u)$

$$\bar{\psi}_s(u) = \frac{1}{\sqrt{s}} \psi\left(\frac{-u}{s}\right). \quad (3)$$

The convolution expression is:

$$Wf(u, s) = (f * \bar{\psi}_s)(u). \quad (4)$$

This can be interpreted as a linear filtering with the impulse response $\bar{\psi}_s(u)$. This equation gives the ability of a filter bank design. A wavelet orthogonal basis can be constructed for discrete parameters $s=2^j$ and $u=i2^j$, named dyadic wavelet transform.

The wavelet transform can be used for different applications. According to Mallat, the wavelet transform, has shown excellent capacities for the detection of signal singularities. When the wavelet function has specific selected properties, wavelet

transform acts as a differential operator. The number of wavelet vanishing moments gives the order of the differentiation. A wavelet $\psi(t)$ is said to have n vanishing moments, when for all positive integer $k < n$, it satisfies the following equation [13]

$$\int_{-\infty}^{+\infty} t^k \psi(t) dt = 0 \quad 0 < k \leq n. \quad (5)$$

A wavelet $\psi(t)$ with a fast decay has n vanishing moments if a function θ has a fast decay such that

$$\psi(t) = (-1)^n \frac{d^n \theta(t)}{dt^n}. \quad (6)$$

This function is called a smoothing function. When the wavelet has n vanishing moments the equation 4 can be written as

$$Wf(u, s) = s^n \frac{d^n}{dx^n} (f * \bar{\theta}_s)(u). \quad (7)$$

Which is a differential operator of order n of the function $f(t)$ smoothed by the function $\bar{\theta}_s$ [14]

$$\bar{\theta}_s = \frac{1}{\sqrt{s}} \theta\left(\frac{-t}{s}\right). \quad (8)$$

Modulus maxima of wavelet coefficients describe the instants such that $|Wf(u, s)|$ is locally maximum.

$$\frac{\partial W f(u, s)}{\partial u} = 0. \quad (9)$$

This equation defines in the time-scale plane, lines corresponding to discontinuities at maximum amplitude of $Wf(u, s)$ [13].

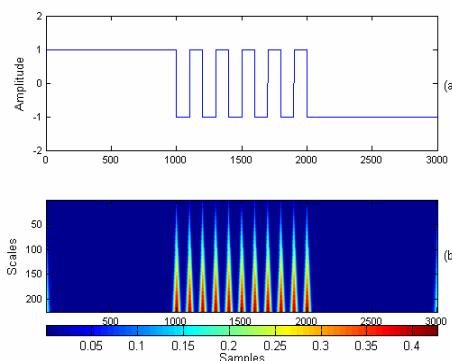


Fig. 1. (a) Square signal and (b) its wavelet transforms at scales ranging from 1/4 to 2

When the wavelet has one vanishing moment, modulus maxima, are the maxima of the first derivative of the smoothed signal at discontinuities [14]. If the wavelet transform has no modulus maxima at fine scales then the function is locally regular. Hence singularities are detected by finding abscissa, where the wavelet modulus maxima converge at fine scales.

Figure 1 gives an example of wavelet modulus maxima amplitude of a square signal. The wavelet used is the quadratic spline wavelet having obviously one vanishing moment.

3 Multi-scale Product for Noisy Signal Edge Detection

Working before the advent of the wavelet framework, Rosenfeld et al suggested forming multi-scale point-wise products [15]. This is intended to enhance multi-scale peaks due to edges, while suppressing noise, by exploiting the multi-scale correlation due to the presence of the desired signal. The multi-scale product (MP) of the dyadic wavelet transform is given by [16], [17]

$$p(u) = \prod_{j=J_0}^J Wf(u, 2^j) . \quad (10)$$

In equation 10, $Wf(u, 2^j)$ represents the wavelet transform at scale $s=2^j$, and the product is operated for scales $2^J, 2^{J-1} \dots 2^{J_0}$. The expression of equation 10 is distinctly a non linear function. The product p reveals peaks at signal edges, and has relatively small values elsewhere. Singularities produce peaks along scale in wavelet transform, these peaks are reinforced by the product $p(n)$. Although particular smoothing levels may not be optimal, the non linear combination tends to reinforce the peaks while suppressing spurious peaks. This is proved by figures 2 and 3.

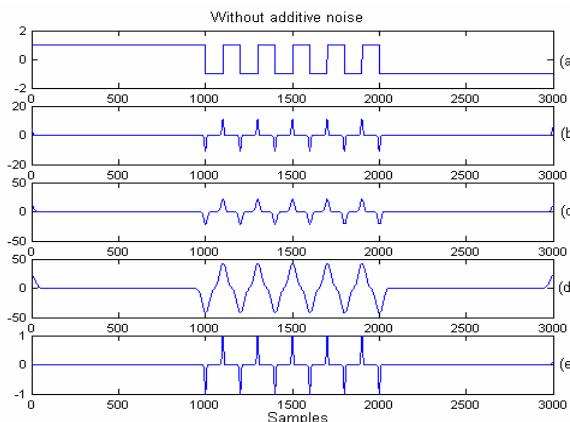


Fig. 2. (a) Square signal, (b) WT at scale $s = \frac{1}{2}$, (c) WT at scale $s=1$, (d) WT at scale $s=2$ and (e) the MP

As illustrative example, figure 2 represents the wavelet transform (WT) of a square signal at the selected scales ($1/2$, 1 and 2) operated with the quadratic spline wavelet, and their products. For this case, the wavelet transform at the lowest scale and the MP permit a better edge localisation.

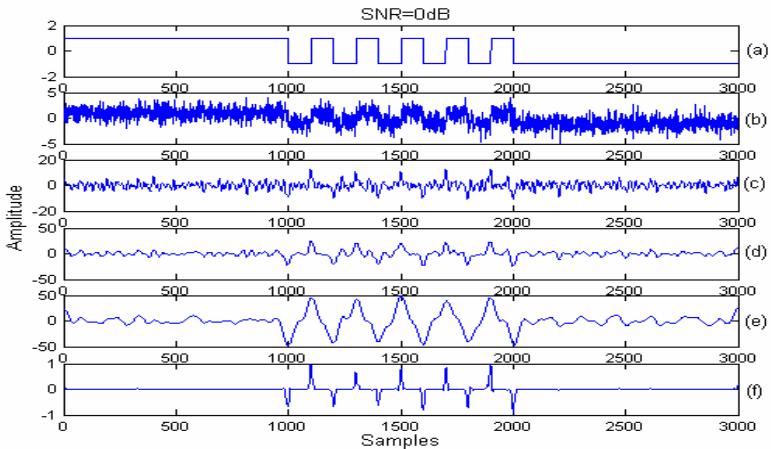


Fig. 3. (a) Representation of the square signal, (b) the signal with additive noise (SNR=0dB), (c) WT at scale $s = 1/2$, (d) WT at scale $s=1$, (e) WT at scale $s=2$ and (f) the MP

Figure 3 illustrates the effect of the wavelet transform and the MP on a signal corrupted with a Gaussian white noise and having an SNR=0dB. The effect of the smoothing due to the scale increasing reduces dramatically the noise power. The noise is completely eradicated by the multi-scale product.

4 Voicing Decision Based on Multi-scale Product

The objective of this section is to illustrate the discontinuity detection capacity of the multi-scale product of speech in presence of powerful additive Gaussian noise. We will show that the multi-scale product reduces additive noise and the intrinsic noise, in other hand it enhances voiced parts of speech and this can be used to operate a voicing classification.

In [18], the multi-scale product is applied on voiced sounds in order to detect speech edges. The cross scale product depicts two types of peaks; minima corresponding to glottal closure instants (GCIs) and maxima related to glottal opening instants (GOIs). In this section, we apply the multi-scale product to a clean and noisy speech signal for voicing detection. The wavelet transforms are calculated at the following successive dyadic scales $1/2$, 1 , 2 corresponding to j values of -1 , 0 and 1 . To show the efficiency of the proposed method, we illustrate in fig. 4(a) the speech signal composed by a silence, voiced and unvoiced sounds corresponding to the segment /and the sun/ extracted from the sentence /the north wind and the sun were disputing which was the stronger/ and uttered by the female speaker f1 of the Keele University database [19].

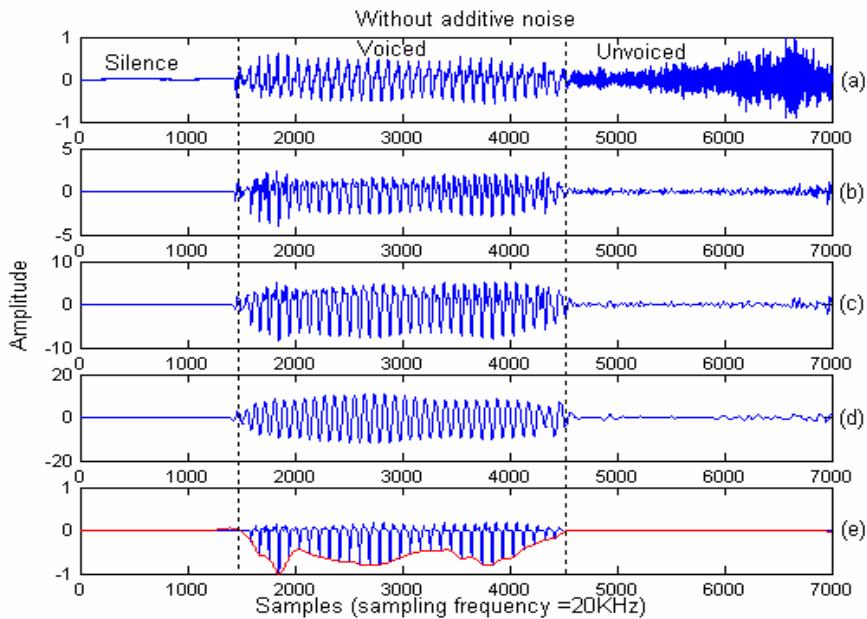


Fig. 4. Voicing decision, (a) speech signal, (b) WT at scale $s=1/2$, (c) WT at scale $s=1$, (d) WT at scale $s=2$ and (e) the product of the three scales giving the detection of the voiced sound

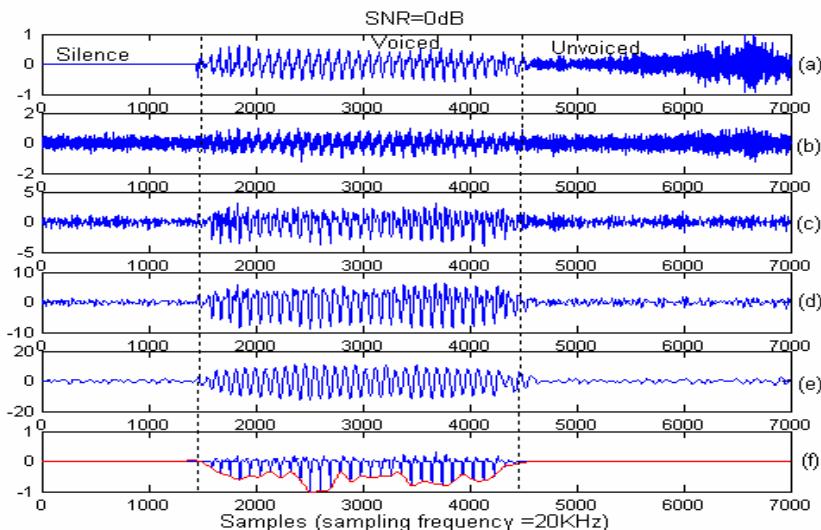


Fig. 5. Voicing decision, (a) speech, (b) speech+ white Gaussian noise SNR=0dB, (c) WT at scale $s=1/2$, (d) WT at scale $s=1$, (e) WT at scale $s=2$, (f) MP

When the scale increases, the unvoiced sounds tend to be annealed by the wavelet transform. The multi-scale product eradicates completely the unvoiced sounds. In fact, in the signal of fig. 4(f), we can only visualize the voiced sound with a periodic structure. This behaviour allows us to make a classification of voiced / (unvoiced sounds or silence). To be able to make a decision, we construct the envelope of minima given by the multi-scale product. This envelope is non null for the voiced sounds. So decision on voicing is made when the corresponding envelope is not annealed.

This approach is also applied to a speech signal corrupted with a Gaussian noise. Figure 5 depicts the speech signal (fig. 5(a)) added to a Gaussian noise with an SNR equal to 0 dB (fig. 5(b)). Its WT at the following scales $\frac{1}{2}$, 1 and 2 are given in fig. 5(c), fig. 5(d) and fig. 5(e). It is shown that at the finest scale, the wavelet coefficients are almost dominated by noise. At the second and third scales, the noise diluted rapidly. It can also be seen that at small scales the positions of the step edges are better localized. But some noise may be falsely considered as edge and the voiced classification can't be operated. At the large scales, the SNR is improved and edges can be detected more correctly but with the decreasing of the accuracy of the edge location.

We can clearly see the effect of the multi-scale product in eliminating the noise and the unvoiced sounds fig. 5(f). The MPM has the same behaviour regarding the noise and the unvoiced sounds.

5 Conclusion

In this work, we have described a new method for voicing decision in the case of noisy speech signal. The proposed approach is based on making the product of the speech signal wavelet transforms at three dyadic scales. This product has an appropriate behaviour regarding the voicing state. The multi-scale product reveals annealed values for silence and unvoiced sounds and has a periodic structure corresponding to voiced sounds. In fact the noise effect is eradicated thanks to the large scale multiplication that reduces the spurious peaks due to noise. This property allows to make a good classification of voiced / (unvoiced or silent), when a noise is added to the speech signal.

References

1. Campbell Jr., J.P.: Speaker Recognition: A Tutorial. Proceedings of the IEEE 85(9), 1437–1462 (1997)
2. Martin, A., Charlet, D., Mauuary, L.: Robust Speech/ Non-speech Detection Using LDA Applied to MFCC. ICASSP, vol. 1, pp. 237-240 (2001)
3. Ishizaka, K., Flanagan, J.L.: Synthesis of voiced Sounds from a Two-mass Model of the Vocal Chords. Bell System Technical J. 50(6), 1233–1268 (1972)
4. Atal, B., Rabiner, L.: A Pattern Recognition Approach to Voiced-unvoiced-silence Classification with Applications to Speech Recognition. IEEE Trans. On Signal Processing 24(3), 201–212 (1976)
5. Kedem, B.: Spectral Analysis and Discrimination by Zero-crossings. In: Proc. IEEE, vol. 74, pp. 1477–1493 (1986)

6. Childers, D.G., Hahn, M., Larar, J.N.: Silence and Voiced/Unvoiced/Mixed Excitation Classification of Speech. *IEEE Trans. On Acoust., Speech, Signal Process* 37(11), 1771–1774 (1989)
7. Liao, L., Gregory, M.: Algorithms for Speech Classification. *5th ISSPA Brisbane*, 623–627 (1999)
8. Zahorian, S.A., Silsbee, P., Wang, X.: Phone Classification with Segmental Features and a Binary-pair Partitioned Neural Network Classifier. In: *ICASSP*, pp. 1011–1014 (1997)
9. Niyogi, P., Sondhi, M.M.: Detecting Stop Consonant in Continuous Speech. *J. Acoust. Soc. Am.* 111, 1063–1076 (2002)
10. Xiong, Z., Huang, T.: Boosting Speech/non-speech Classification Using Averaged Mel-frequency Cepstrum. In: *IEEE Pacific-Rim Conf. on Multimedia* (2002)
11. Yang, H., Vuuren, S.V., Hermansky, H.: Relevancy of Time-frequency Features for Phonetic Classification Measured by Mutual Information. In: *ICASSP*, vol. 1, pp. 225–228 (1999)
12. Lachiri, Z., Ellouze, N.: Speech Classification in Noisy Environment Using Subband Decomposition. In: *ISSPA*, vol. 1, pp. 409–412 (2003)
13. Mallat, S.: *A Wavelet Tour of Signal Processing*, 2nd edn. Academic Press, San Diego (1999)
14. Mallat, S., Hwang, W.L.: Singularity Detection and Processing with Wavelets. *IEEE Trans. On Information Theory* 38(2), 617–643 (1992)
15. Rosenfeld, A., Thurson, M.: Edge and Curve Detection for Visual Scene Analysis. *IEEE Trans. Comput.* 20, 562–569 (1971)
16. Sadler, B.M., Pham, T., Sadler, L.C.: Optimal and Wavelet-based Shock Wave Detection and Estimation. *Journal of the Acoustical Society of America* 104(2), 955–963 (1998)
17. Sadler, B.M., Swami, A.: Analysis of Multiscale Products for Step Detection and Estimation. *IEEE Trans. Inform. Theory* 45(3), 1043–1051 (1999)
18. Bouzid, A., Ellouze, N.: Open Quotient Measurements Based on Multiscale Product of Speech Signal Wavelet Transform. *Research Letter in Signal Processing 2007*, p. 5 (2007); Article ID 62521 doi:10.1155/2007/62521
19. Plante, F., Meyer, G.F., Ainsworth, W.A.: A Pitch Extraction Reference Database. In: *Proc. Eurospeech 1995*, pp. 837–840 (1995)

Coherent Synchronization of a Chaotic Communication System Based on a QAM-4 Transmitter

Roxana Burgelea¹, Jean Philippe Mangeot¹, Frédéric Launay¹,
Patrick Coirault¹, and Safwan El Assad²

¹ LAII - ESIP, Poitiers, France

frederic.launay@univ-poitiers.fr,

<http://laii.univ-poitiers.fr>

² Polytech'Nantes, Laboratoire IREENA, Batiment IRESTE
F-44000 Nantes, France

Abstract. Remarkable research efforts have been invested in recent years on the chaotic communications schemes. In this paper we propose a modulation technique based on QAM-4 (Quadrature Amplitude Modulation). The data is split into I and Q channels. Each channel is spread by chaotic sequences generated over a fixed time interval by a specific digital chaotic system. There are two principal advantages of introducing the chaos in a communication scheme. The first is that it allows signals to be transmitted in a secure way, so that the signal could be perceived as a noise for a potential intruder. Secondly, the spread-spectrum characteristics of the signal improve the noise rejection properties. A new coherent receiver structure is presented to synchronize each channel. Simulation results are presented to show the effectiveness of the proposed scheme.

1 Introduction

Nowadays, a great deal of research is focused on chaotic communications schemes. The potential benefits that can be gained from using chaotic signals, including robustness in multipath environments, ease of spectrum spreading, added security etc has made them promising for communications systems within various sectors on industry [1]. In this paper we propose an enhanced QAM-4 (or QPSK - Quadrature Phase Shift Keying) scheme where a chaotic encoder spreads the input sequence. The data is split into two channels and transmitted from one location to another by mapping bit sequences to symbols, and symbols to sample functions of chaotic waveforms (generated over a fixed time interval by digital chaotic systems [2]). The signals are then filtered and modulated around two quadrature carriers according to the QAM modulation technique. The resulted analog waveform passes through a band limited analog channel, where the signal is distorted and noise is added.

At the receiver, the encoded information can be extracted efficiently by means of coherent detection, where all possible sample functions are known, or by non-coherent detection, where one or more characteristics of the sample functions are

estimated. First of all, the received quadrature carriers are downconverted and filtered. Furthermore the data is limited at its original spectrum by multiplying it with the original chaotic sequences and finally it is combined into the original modulating information. The proposed receiver structure is based on a coherent method. The article provides in brief the advantages and disadvantages of such a system compared with non coherent method.

The contribution of this paper is to define the characteristics of the chaotic generator in order to improve error detection, and then to propose an algorithm to correct and consequently to synchronize the receiver in a QAM scheme.

This paper is organized as follows: in Section II we focus on the conception of the QAM-4 chaotic modem. In Section III a coherent demodulation technique is proposed with a specific chaotic generator. In Section IV numerical results are provided to illustrate the effectiveness and the efficiency of the proposed scheme. Conclusions and suggestions for further works are presented in Section V.

2 Conception of a QAM Chaotic Modem [2]

The QAM-4 transmitter based on chaotic spread spectrum signals is composed of a QAM coder which splits the binary sequence $b(t)$ into I and Q channels. The transmitted sequence $b(t)$ is a periodic binary sequence of T_b period (let's rect_{T_b} a rectangle function)

$$b(t) = \sum_{k=1}^{\infty} b_k \cdot \text{rect}_{T_b}(t - kT_b) \text{ with } b_k \in \{0, 1\} \quad (1)$$

Each channel is spread out by a continuous chaotic code. The chaotic generator delivers a data sequence whose value is included in the integer interval [0,254] which is converted to 8 serial bits. The coding operation is carried out by inserting the code or the opposite of the code depending if a +1 or a -1 is transmitted.

In order to product spread spectrum signal, chaotic sequences (characterized by a higher rate than that one of the symbol's component) are introduced for each symbol period as specified in the introduction part. The chaotic I and Q coded sequences can be mathematically expressed as:

$$c_{i,q}(t) = \sum_{k=1}^{\infty} c_{i,k}^{N_{ch}} \cdot \text{rect}_{T_{ch}}(t - kT_{ch}) \quad (2)$$

where T_{ch} represents the sequence period and $c_{i,k}^{N_{ch}}, c_{i,k}^{N_{ch}}$ are two vectors of integer values included in the [0,254] interval. Finally, the decimal chaotic sequence is converted into an 8-bit binary code by a Parallel/Series block. Another one chaotic generator based technique was studied by Galias [3]. Each channel is coded by two quadrature chaotic sequences obtained from a Hilbert filter at the output of the chaotic generator.

Then, the signals are filtered through a raised cosine Nyquist filter. After filtering operation, filtered signals are multiplied around two quadrature carriers characterized by the F_c frequency.

The figure 1 represents the general diagram of the chaotic transmitter.

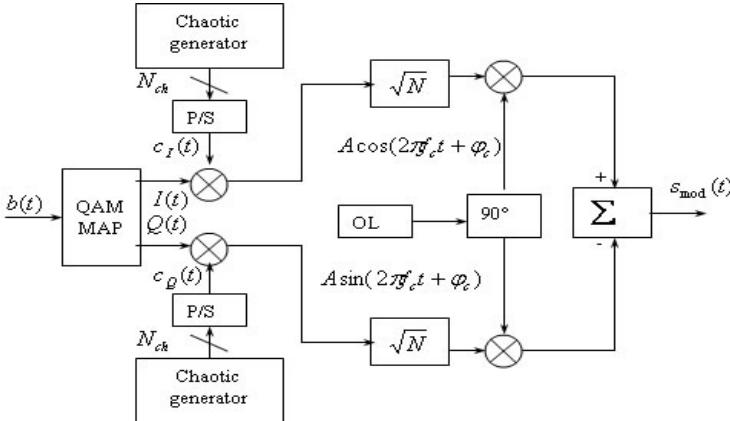


Fig. 1. QAM chaotic transmitter

At the reception, two operations are carried out : The modulated signal is filtered and brought back into the original base band.

If the chaotic coded sequence is known by the receiver, this base band signal is then filtered and de-spread by correlation operation in order to rebuild the original I and Q sequences. Finally, a level detection followed with a re-arrangement of the I and Q data entries makes it possible to recover the initial sequence $b(t)$. However, due to the sensitivity of chaotic systems to initials conditions, permanent synchronicity between receiver and transmitter is necessary or else the trajectories of the two generators will be quickly uncorrelated. Moreover, the receiver must be able to synchronize itself on the associated generator and that in spite of the noise in the channel. A robust method to re-build the chaotics sequences is the crucial point.

Several receiver structures for chaotic spreaded system is based on a DCSK non-coherent method [5]. In this case, every transmitted symbol is represented by a pair of chaotic signal samples sent in two equal time slots:

- the chaotic code sent in the first half of the symbol period serves as the reference (reference sample).
- the reference code or its inverted version sent in the second half of the symbol period carries the information and represents the data sample.

Such structure can be easily implemented for QAM chaotic systems. Since the reference code is sent with the modulated data, permanent synchronization of the receiver is assured. It's the main advantage because two chaotics systems

diverge quickly even if initials conditions are very close : As coded states are transmitted, the receiver does not try to synchronize its chaotic sequence from the corrupted sequence but only uses a delay to correlate modulated code with the reference one. Nevertheless since no modulated chaotic sequence is clearly sent, we can not prevent intruder to recover transmitted signal unless this one is sent in different carriers : coherent method present a better secure way.

3 Coder Demodulation Technique for a QAM Chaotic System

In this section we present a method to build a chaotic generator. The generator is based on Frey model [4] but in our model the coefficients are based on maximal Hamming distance between the transmitted code and the estimated one when no more one error occurred in each previous codes. Nevertheless optimal generator in the sense of Hamming may yield to a poor cyclic sequence. As a result, one may prefer degrading the Hamming distance to improve the chaotic trajectory of the generator.

With Hamming distance, the demodulation technique proposed is based on statistical properties of chaotic sequence in a bounded noise. This method consists in estimating the next chaotic code according to the previous state in presence of limited noise.

The chaotic sequence c_k is defined by the previous code (at least c_{k-1} and c_{k-2}) and by a non linear function. By knowing the coder function, and thanks to the first chaotic codes, the chaotic sequence \hat{c}_k can be rebuilt. Nevertheless, each code of the received chaotic sequence is corrupted by an additive Gaussian white noise. As a result, estimation of next code lead to several set of possible state named \hat{S} and can quickly diverge to all possible state.

Once Hamming distance is optimal (by choosing the right characteristics of chaotic generator), we propose an algorithm based on correlation between estimated state from interval of error of previous state and noisy last received states: With c_{k-1} and c_{k-2} and bounded relative error, we can remove some state of \hat{S} in order to reduce the set of possible states.

3.1 Definition of Chaotic Generator Characteristics

First to improve the receiver performance, we aim to design a chaotic modulator according to some assumptions.

Each code is defined over 8 bits. Let's assume that no more one error is occurred by code. As a result, we aim to find the best coder function to detect and correct the wrong bit.

To improve detection of error and assuming that no more one error can occur in each code, we propose a coder function which Hamming distance is maximal between each error code. As one error can occur in the last code and in the previous one, the altered estimation code in the receiver from the last two codes should to be the more distant compared with the received code.

As a result, the coder function is defined by the following chaotic function :

$$\begin{cases} x[k] = \text{modulo}(G_1.x[k-1] + G_2.x[k-2] + s_{k-1}, 255) \\ G_1 = 9 \\ G_2 = 6 \\ s_{k-1} = 3 \text{ if } x_{k-1} > 127 \end{cases} \quad (3)$$

When two codes are received, an algorithm assessed the third code. This estimated code, named \hat{c}_k is compared to the transmitted one.

For example, if the first code is 4 and second code is 1 then the third code, according to (eq 7) is 33. The binary transmitted sequence is (4) -00000100, (1) - 00000001, (33) - 00100001. If first code received is (36) - 00100100, the second one is (3) 00000011, then the estimated code is $\hat{c}_k = 11110011$. Thus, compared to (33) - 00100001, the hamming distance is equal to 4 as we can find in table at line 7 (bit 6 of first code is false), and at column 3 (bit 2 of last code is false)

In a general case, let's $x[k-2]$ and $x[k-1]$ the transmitted code. We suppose that an error occurred on $x[k-1]$ and $x[k-2]$. To simplify, suppose that an error transforms a '0' to '1'. The received codes are $x_r[k-2] = x[k-2] + 2^i$ and $x_r[k-1] = x[k-1] + 2^j$ where i and j represent the position of the error. The estimated code is : $\hat{x}[k] = \text{modulo}(G_1.(x[k-1] + 2^j) + G_2.(x[k-2] + 2^i) + \hat{s}_{k-1}, 255)$

We estimate the error between $x[k]$ and $\hat{x}[k]$ thus $E[k] = \hat{x}[k] - x[k]$

$$\begin{aligned} E[k] &= \text{modulo}(G_1.(x[k-1] + 2^j) + G_2.(x[k-2] + 2^i) + \hat{s}_{k-1}, 255) \\ &\quad - \text{modulo}(G_1.x[k-1] + G_2.x[k-2] + s_{k-1}, 255) \end{aligned} \quad (4)$$

with $x[k]$ and $\hat{x}[k]$ comprised between (0,254).

As a result, 4 can be written as

$$\begin{aligned} E[k] &= (G_1.(x[k-1] + 2^j) + G_2.(x[k-2] + 2^i) + \hat{s}_{k-1} - 255.\alpha) \\ &\quad - (G_1.x[k-1] + G_2.x[k-2] + s_{k-1} - 255.\beta) \end{aligned} \quad (5)$$

Finally, with modulo 255, we can write :

$$\begin{cases} E[k] = \text{modulo}(G_1.2^j + G_2.2^i + \Delta s_{k-1}, 255) \text{ if } \hat{x}[k] > x[k] \\ E[k] = 255 - \text{modulo}(G_1.2^j + G_2.2^i + \Delta s_{k-1}, 255) \text{ if } \hat{x}[k] < x[k] \end{cases} \quad (6)$$

The following table represents the Hamming distance of transmitted and estimated codes when an error occurred both in the first code and in the second

one and when $\hat{x}[k] > x[k]$. The lines represent the error position of the first code and the columns represent the error position of the second code (from the least significant bit -1- to the most significant bit -8-). Of course, when no error occurred (first line, first column), Hamming distance is null.

Table 1. Table of Hamming distance when $\hat{x}[k] > x[k]$

	0	1	2	3	4	5	6	7	8
0	0	2	2	2	2	2	2	2	4
1	2	4	2	3	4	4	4	2	4
2	2	3	4	2	3	4	4	4	4
3	2	2	3	4	2	3	4	4	6
4	2	4	2	3	4	2	3	4	6
5	2	4	4	2	3	4	2	3	6
6	2	4	4	4	2	3	4	2	2
7	2	3	4	4	4	2	3	4	2
8	2	2	3	4	4	4	2	3	3

	0	1	2	3	4	5	6	7	8
0	0	6	6	6	6	6	6	6	6
1	2	6	6	4	6	5	4	4	6
2	2	2	6	6	4	6	5	4	2
3	2	4	2	6	6	4	6	5	2
4	2	4	4	2	6	6	4	6	3
5	2	5	4	4	2	6	6	4	4
6	2	6	5	4	4	2	6	6	4
7	2	4	6	5	4	4	2	6	6
8	2	6	4	6	5	4	4	2	6

	0	1	2	3	4	5	6	7	8
0	0	6	6	6	6	6	6	6	6
1	6	4	6	5	4	4	4	6	4
2	6	5	4	6	5	4	4	4	4
3	6	6	5	4	6	5	4	4	2
4	6	4	6	5	4	6	5	4	2
5	6	4	4	6	5	4	6	5	2
6	6	4	4	4	6	5	4	6	6
7	6	5	4	4	4	6	5	4	6
8	6	5	4	4	4	6	5	5	5

(a)

(b)

(c)

(d)

(a) when '0' become '1' of both code, (b) when '1' of $x[k-1]$ become '0' and '0' of $x[k-2]$ become '1' (c)when '0' become '1' of $x[k-1]$ and '1' become '0' of $x[k-2]$ and (d) when '1' become '0' of $x[k-1]$ and '1' become '0' of $x[k-2]$

According to such a function, the Hamming error is maximal between the received code and the estimated code when one error occurred in the first and/or in the second code. As a result assuming one error in the first and/or in the second code, the Hamming distance between the transmitted and estimated sequence is at least equal to 2. Thus, even if an error occurred in the third code, the Hamming distance will be greater than 1.

Modulo 255 is preferred to 256, since with modulo 256 no error of the estimated code compared with the received code is generated when the 7th bit of first and the 6th bit of the second code are '1' instead of '0' (and according with the generator characteristics : $\text{modulo}(2^7 * G_1, 2^6 * G_2, 256) = 0$).

More generally, $\text{modulo}(2^7 * G_1, 2^7 * G_1, 256)$ is null if G_1 and G_2 are both odd. Finally, all combinaisons of G_1 and G_2 with one error of the two codes can lead to an estimated code which is equal to the received code with modulo 256.

Moreover, the Hamming distance on the previous tables is estimated when $\hat{x}[k] > x[k]$. Thanks to modulo 255, if $\hat{x}[k] < x[k]$, the Hamming distance is the complementary. As an example if Hamming distance is 3 when $\hat{x}[k] > x[k]$ (three bits over 8 are different) so, the Hamming distance becomes 5 if $\hat{x}[k] < x[k]$. As Hamming distance in tables 1-a to 1-d are comprised between 2 and 6 when $\hat{x}[k] > x[k]$, thus if $\hat{x}[k] < x[k]$ the Hamming distance are also comprised between 2 and 6. It's an important result since if no more one error occurred in the first code and in the second code, the Hamming distance between the

estimated code and the third received code (with no more one error) will never be null.

3.2 Error Correction to Synchronize the Receiver

Let's c_k the integer chaotic code defined between $[0,254]$, c_k^t the binary transmitted code and c_k^r the binary received code. If information is '0' the transmitted code is the complement of the binary chaotic code. As a result, the integer chaotic transmitted code $c_k^t = c_k$ if '1' is sent or $c_k^t = \text{modulo} - c_k$ if '0' is sent. In the first step, we assume a transmission without error.

According (7), the estimated code \hat{c}_k is $G_2.c_{k-2}^r + G_1.c_{k-1}^r + s_{k-1}$ if (1,1) is transmitted, $G_2.c_{k-2}^r + G_1.(\text{modulo} - c_{k-1}^r) + s_{k-1}$ if (1,0) is transmitted, ...

First, note that

$$\left\{ \begin{array}{l} G_2.(\text{modulo} - c_{k-2}^r) + G_1.(\text{modulo} - c_{k-1}^r) = \text{modulo} - (G_2.c_{k-2}^r + G_1.c_{k-1}^r) \\ G_2.(\text{modulo} - c_{k-2}^r) + G_1.c_{k-1}^r = \text{modulo} - (G_2.c_{k-2}^r + G_1.(\text{modulo} - c_{k-1}^r)) \\ \text{if } c_k < 128 \text{ then } \text{modulo} - c_k \text{ is greater than 128.} \end{array} \right. \quad (7)$$

With this last condition, there exists one solution which is equal to c_k or $(\text{modulo} - c_k)$ and as a result an unique 3-plet information bits M_{k-2}, M_{k-1}, M_k .

To determine if an error occurred, we have to compare c_k with \hat{c}_k using the xor function. If there are no difference, the information bit M_k is '1', if all bits are false then there are no error and the information bit M_k is '0'.

When no more than one error occurred in the first and second code, and according to the 4 possible states of information (00, 01, 10, 11) we extract from the table 1 the possible combinations of code c_{k-2}^r and c_{k-1}^r which can lead to this difference.

As a result, we define the following algorithm to correct the bit error.

Algorithm

First Step : First codes estimation

1. We assess \hat{c}_k with c_{k-2}^r and c_{k-1}^r .
2. We compare \hat{c}_k with c_k^r (XOR function). The number of errors E_k gives us information about where the error may have occurred.
3. Comparison with table 1 if no error occurs in c_k^r (3-1) Assuming that $M_{k-2} = 1$ and $M_{k-1} = 1$, and with Hamming distance E_k or $255 - E_k$ according to the comparison of \hat{c}_k and c_k^r , we reduce the possible states \hat{c}_{k-2} and \hat{c}_{k-1} from the table 1 to $\hat{S}_{k-2}^{1,1}$. According to the Table 1a, we only save the states which lead to a number of errors E_k . The table 1a represents the position of the error which transforms a '0' to '1', the line is the first code and the column the second code. Let's $P1(k1,k2)$ the position of the table 1a where Hamming distance is E_k . $P1(k1,k2)$ represents the position of a possible error of c_{k-2}^r

and c_{k-1}^r where a '0' was transformed to a '1'. Nevertheless, suppose that at this bit position, the received bit (either on c_{k-2}^r or on c_{k-1}^r) is equal to '0', thus this case is not a feasible error state. As a result this second test reduces all the possible cases : if the k1 bit of c_{k-2}^r is '1' AND the k2 bit of c_{k-2}^r is '1', then we extract and save a possible right code.

Same procedure with table 1b, 1c and 1d 3-2) Assuming that $M_{k-2} = 1$ and $M_{k-1} = -1$, and E_k is the Hamming distance between c_k and \hat{c}_k . The possible state of error from \hat{c}_{k-2} and $modulo - \hat{c}_{k-1}$ are reduced from the table 1 to $\hat{S}^{(k-2)}_{1,0}$ (taking into account all possible errors ie, '0' become '1' and '1' became '0' and noting that if an error transform a bit '0' of $modulo - \hat{c}_{k-1}$ to '1' then the table of reference correspond to a '1' that becomes '0').

4. Coming back from point 2 to point 4 assuming an error on c_k^r In this state, we are compelled to take into account the error on c_k^r to estimate c_k^c , the estimated corrected code. Then, according to \hat{c}_k is greater than c_k^c or not, the Hamming distance is E_k or $8 - E_k$. All possible state of $M_{k-2} = 1$ and $M_{k-1} = -1$ have to be tested.

As a result, we find some combinaison of u-plet $\hat{c}_{k-2}, \hat{c}_{k-1}, \hat{c}_k$ with associated bit information M_{k-2}, M_{k-1}, M_k . We save these u-plet in a table T_k .

Second Step : Convergence

Come back from Step 1, with c_{k-1}^r instead of c_{k-2}^r and c_k^r instead of c_{k-1}^r . We save the u-plet in a Table T_{k+1} .

The convergence is given by comparing the possible code \hat{c}_{k-1}, \hat{c}_k with information M_{k-1}, M_k of T_{k+1} and the same possible code saved at T_k .

According to each state \hat{S}^k and \hat{c}_k , we estimate \hat{c}_{k+1} . Nevertheless, \hat{c}_{k+1} is generated from the second bit of \hat{S}^k and only the 8 possible states of \hat{c}_{k+1} . As a result, among \hat{S}^k couples of code, each code of the second estimated code \hat{c}_k is tested with c_{k+1} .

4 Numerical Results

Simulations have been made when no information was sent (M_k). To prove the effectiveness of our algorithm, we generated three positive consecutive errors ('0' became '1'). As a result, the number of possible case of errors is given by the number of Table 1a equal to Hamming distance, Hamming distance -1 and Hamming distance +1.

Indeed, let's the following chaotic sequence 4,1,33, 48. An error is added to the first code and to the second code at the 4th bit and in the third code to the 6th bit. The first three received code are 12, 9 and 97. According to the two first received code (12,9), the estimated third code is 153 (1 0 0 1 1 0 0 1). Compared to 97 (01100001), there are 5 errors. Nevertheless, as an error may have occurred in the third code, we are compelled to test three Hamming distance : 4, 5 and 6.

The total number of possible error is 34. Moreover, considering that probability of Hamming distance equal to 4 is greater than Hamming distance is equal

to 5, then we can conclude that there are a greater probability that an error occurred in the third received code.

Nevertheless, thanks to the third point of the algorithm (First Step) we can drastically reduce the number of combinaisons. Indeed, according to table 1a, a distance of Hamming equal to 3 can be generated by an error ('0' become '1' in the third bit of the first code and in the second bit of the second code. In our exemple, the binary value of 12 is 00001100 and the binary value of 9 is 00001001. The third bit of the first code is 1 but the second bit of the second code is 0. So, since the second bit is not equal to 1 (condition of Hamming error of Three) this couple of point can not be a possibility.

As a result, we only find 5 possible errors at step 1. Let's T1 the table where the 3-plet of 5 estimated chaotic code after correction are saved $\hat{c}_1, \hat{c}_2, \hat{c}_3$.

At step 2, the chaotic sequence is shifted of one bit and launch the first step of the algorithm. We only find 6 possibles chaotic sequence after correction. Let's T2 the table of possible corrected codes $\hat{c}_2, \hat{c}_3, \hat{c}_4$.

By comparison of \hat{c}_2, \hat{c}_3 in T1 and T2, we can estimate/correct and recover the chaotic sequence of the transmitter.

As a result, with no more than one error per byte and no modulated data, convergence is achieved with a few steps. nevertheless, with modulated data, more memory is necessary to test and save several ways.

5 Conclusion

This paper deals with a new demodulation technique based on QPSK. This schema is interesting for two reasons: on one hand the spread-spectrum characteristics of the signals authorize the usage in a multi-user system such as CDMA ([6], [7]), on the other hand it makes it possible to decrease the error rate. In addition, the proposed schema does not require synchronization between the receiver and the transmitter, which constitutes a great advantage compared to the coherent detectors, especially knowing that the synchronization becomes more difficult in noise presence. Since the system is implemented in a full digital way, it makes it possible to control efficiently control all of the coder parameters. A digital signal has exact characteristics and it is possible to ensure a better interfacing with a computer data. Moreover synchronization is achieved by correcting the potential error. To prevent burst of errors we also can interleave the chaotic bits.

Nevertheless, this approach is still limited in a low noise environment, since no more than one error is accepted per code.

References

1. Lau, F.C.M., Tse, C.K.: Chaos Based Digital Communication Systems (2002)
2. Assad, S.E., Vladeanu, C.: Digital chaotic codec for DS-CDMA Communication Systems. Lebanese Science Journal 7(2), 55–71 (2006)
3. Galias, Z., Maggio, G.M.: Quadrature Chaos Shift Keying: Theory and Performance Analysis. IEEE Transaction on Circuits and Systems 48(12) (2001)

4. Frey, D.R.: Chaotic Digital Encoding: An Approach to Secure Communication. *IEEE Trans. Circuits and Syst. II: Analog and Digital Signal Processing* 40(10), 660–666 (1993)
5. Kolumban, G., Vizvari, B., Schwarz, W., Abel, A.: Differential Chaos Shift Keying: A robust coding for chaos communication, Seville, Espagne, june 27-28 (1996)
6. Yang, T., Chua, L.O.: Chaotic Digital Code-Division Multiple acces (CDMA) Communication Systems. *International Journal of Bifurcation and Chaos* 7(12) (1997)
7. Moshavi, S., Bellcore: Multi-user detection for DS-CDMA communications. *IEEE Communications Magazine* (October 1996)

Joint Multiple Target Tracking and Classification Using Controlled Based Cheap JPDA-Multiple Model Particle Filter in Cluttered Environment

Zahir Messaoudi¹, Abdelaziz Ouldali¹, and Mourad Oussalah²

¹ Military Polytechnic School, BP 17 Bordj El Bahri Algiers, Algeria

² University of Birmingham, Electronics, Electrical and Computer Engineering
Edgbaston, Birmingham B15 2TT, UK
messaoudi_zahir06@yahoo.fr

Abstract. In this paper, we address the problem of jointly tracking and classifying several targets in cluttered environment. It is assumed that the motion model of a given target belongs to one of several classes. We propose to use the multiple model particle filter (MMPF) to perform nonlinear filtering with switching dynamic models. Moreover, the principle of joint probabilistic data association (JPDA) is used to determine the measurements origin. Besides, the joint probabilities are calculated using Fitzgerald's had hoc formulation (Cheap JPDA) whose efficiency has been proven in the literature. On the other hand, a controller based on the quality of the innovation has been implemented in order to tune the number of particles. The feasibility and the performances of the proposal have been demonstrated using a set of Monte Carlo simulations dealing with two maneuvering targets.

1 Introduction

Tracking multiple targets arises in many engineering applications where a discrete uncertainty regarding measurement-target associations takes place. In radar tracking it is uncertain whether the radar echo stands for a friendly or an enemy aircraft or just a false alarm, whether it is a commercial or a military transport aircraft [1]. To solve the problem of data association, the joint probabilistic data association filter (JPDAF) has been successfully applied in many tracking examples [2,3]. The computational complexity for the joint probabilities increases exponentially as the number of targets increases. To reduce this computational complexity, Fitzgerald [8,9] developed the simplified version of the JPDAF, called the cheap JPDAF (CJPDAF) algorithm. However, the nonlinearities in the target/measurement, especially for highly manoeuvring targets, have limited the use of the JPDAF. These nonlinearities have often been addressed through the linearization around the state estimates. Unfortunately, the performance of the state estimation degrades as the non-linearities become more severe [4]. With, the particle filter (PF) [5-7] it is possible to overcome the effect of nonlinearity as well as the standard Gaussian assumption. Moreover, the multiple-model PF (MMFP)

can be used to perform nonlinear filtering with switching dynamic models using an augmented state vector which consists of both target kinematics variables and discrete-valued part (regime variable).

In this paper, we address the problem of jointly tracking and classifying several targets using CJPDA-MMPF in cluttered environment. The process of jointly tracking and classifying is based on the following ideas: i) The estimation process (target kinematics variables and regime variable) is carried out at each particle level, ii) The calculus of the joint probabilities is accomplished using CJPDAF, iii) Similarly to [9], a controller is used to tune the number of controller at each iterations. Section 2 of this paper presents some aspects of the JPDAF and relates Fitzgerald's had hoc formulation. Section 3 presents the multiple switching dynamic model. In section 4, we propose the CJPDA-MMPF. In section 5, some simulation examples illustrate the feasibility and performances of the proposal.

2 Data Association Method

One of the most popular method for data association is probably the JPDAF. A detailed derivation of JPDAF can be found in [2,3]. The Kalman update equation is (one uses superscript t to refer to target t)

$$\hat{x}^t(k) = \hat{x}_p^t(k) + \mathbf{K}^t(k)v^t(k) \quad (1)$$

where $\hat{x}_p^t(k)$ is the predicted state vector, $\mathbf{K}^t(k)$ is the Kalman gain and $v^t(k)$ is the combined innovation given by

$$v^t(k) = \sum_{i=1}^{m^t(k)} \beta_i^t(k)v_i^t(k) \quad (2)$$

where $\beta_i^t(k)$ is the probability of associating track t with measurement i , $m^t(k)$ is the number of validated measurements for track t and $v_i^t(k)$ is the innovation of track t and measurement i .

In the standard JPDAF, the association probabilities $\beta_i^t(k)$ are calculated by considering every possible hypothesis as to the association of the new measurements with existing tracks. A had hoc JPDAF formulation to get very easily $\beta_j^t(k)$'s values was pointed out by Fitzgerald [8,9] as

$$\beta_i^t(k) = G_i^t(k)[S^t(k) + S_i(k) - G_i^t(k) + B]^{-1} \quad (3)$$

where $G_i^t(k)$ is the distribution of $v_i^t(k)$ usually assumed to be gaussian and

$$S^t(k) = \sum_{i=1}^{m^t(k)} G_i^t(k) \quad , \quad S_i(k) = \sum_{t=1}^M G_i^t(k)$$

with M is the number of targets and B is a constant which depends on clutter density (usually with $B = 0$, the algorithm works well).

3 Multiple Switching Dynamic Model

In many engineering applications one deals with nonlinear dynamic systems characterized by some modes of operation. For example, this is the case of maneuvering targets. These kinds of problems are referred to as hybrid state estimation problems involving both target state and mode variable. Generally, an hybrid system is described by the following equations [10]

$$x(k+1) = f(x(k), r(k), u(k)) \quad (4)$$

$$z(k) = h(x(k), r(k), w(k)) \quad (5)$$

where $r(k)$ is the mode variable commonly modeled by time-homogeneous s -state first-order Markov chain with transitional probabilities

$$\Pi_{i,j} = \text{Prob}(r(k) = j / r(k-1) = i) \quad (i, j \in \{1, 2, \dots, s\}) \quad (6)$$

In this work, two different motion models are considered for target t : the constant velocity and the coordinated turn rate models.

Constant velocity model: The target t is assumed to move with a constant velocity. For notational simplicity, $x^t(k)$ refers to the state (coordinates and velocities) of the target following this motion. We denote α^t and ζ^t the coordinates: $x^t(k) = [x_1^t(k), x_2^t(k), x_3^t(k), x_4^t(k)]^T = [\alpha^t, \dot{\alpha}^t, \zeta^t, \dot{\zeta}^t]^T$. In the following, Δ is assumed to be the sampling period.

$$x^t(k+1) = \mathbf{F}^t(r^t(k) = 1)x^t(k) + \mathbf{D}u(k) = \begin{bmatrix} 1 & \Delta & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \Delta \\ 0 & 0 & 0 & 1 \end{bmatrix} x^t(k) + \mathbf{D}u(k) \quad (7)$$

where $\mathbf{D} = \begin{bmatrix} \frac{\Delta^2}{2} & \Delta & 0 & 0 \\ 0 & 0 & \frac{\Delta^2}{2} & \Delta \end{bmatrix}^T$ and $u(k)$ is white noise ($u(k) \sim \mathcal{N}(0, Q)$).

Coordinated turn rate model: In this model, the target t moves with a constant velocity and a constant known turn rate ω^t . Again $x^t(k)$ refers to the state (coordinates and velocities) of the target following this motion. To obtain this model, we only change the transition matrix $\mathbf{F}^t(r^t(k) = 1)$ in (7) by $\mathbf{F}^t(r^t(k) = 2)$ given by

$$\mathbf{F}^t(r^t(k) = 2) = \begin{bmatrix} 1 & \sin(\omega^t \Delta) & 0 & \frac{(1-\cos(\omega^t \Delta))}{\omega^t} \\ 0 & \cos(\omega^t \Delta) & 0 & -\sin(\omega^t \Delta) \\ 0 & \frac{-(1-\cos(\omega^t \Delta))}{\omega^t} & 1 & \frac{\sin(\omega^t \Delta)}{\omega^t} \\ 0 & \sin(\omega^t \Delta) & 0 & \cos(\omega^t \Delta) \end{bmatrix} \quad (8)$$

Measurement equation: The measurements are modeled by

$$z^t(k) = \phi(x^t(k)) + w(k) = \begin{bmatrix} \sqrt{(x_1^t)^2(k) + (x_3^t)^2(k)} \\ \arctan\left(\frac{x_3^t(k)}{x_1^t(k)}\right) \end{bmatrix} + w(k) \quad (9)$$

where $w(k)$ is white noise ($w(k) \sim \mathcal{N}(0, R)$).

4 CJPDA-MMPF

The proposal developed in this paper is based on the following principle : i) The estimation process of the augmented state vector (target state and mode variable) is accomplished at each particle level, ii) The CJDAF is used in order to estimate the joint probability association target / measurement, iii) Similarly to [9], a classic controller is used to control the number of particles.

A pseudo code of a generic CJPDA-MMPF is mentionned below.

1. Initialization: Set $k = 0$
 - Generate N_p samples $\{\hat{x}_n^t(0)\}_{n=1}^{N_p}$ for all targets from $p(x^t(0))$
 - Generate N_p samples $\{r_n^t(1)\}_{n=1}^{N_p}$ for all targets from $p(r^t(1))$
2. Increase k
3. For each target t ($t = 1, \dots, M$) do
 - Generate a random set $\{r_n^t(k)\}_{n=1}^{N_p}$ based on $\{r_n^t(k-1)\}_{n=1}^{N_p}$ and the transition probabilities matrix $\mathbf{\Pi}$ (see table 3.9 [10]).
 - for each particle n ($n = 1, \dots, N_p$) do
 - Compute $\{\beta_i^t\}_{i=1}^{m^t(k)}$ according to (3) based on $r_n^t(k), x_n^t(k-1)$ and validated measurements $\{z^i(k)\}_{i=1}^{m^t(k)}$
 - Evaluate

$$\hat{x}_{p,n}^t(k) = \mathbf{F}^t(r_n^t(k)) \hat{x}_n^t(k-1)$$

$$\hat{\mathbf{P}}_{p,n}^t(k) = \mathbf{F}^t(r_n^t(k)) \hat{\mathbf{P}}_n^t(k-1) [\mathbf{F}^t(r_n^t(k))]^T + \mathbf{D} \mathbf{Q} \mathbf{D}^T$$

$$\mathbf{H}(\hat{x}_{p,n}^t(k)) = \text{Jacobeann, of } \phi \text{ given in (9), evaluated at } \hat{x}_{p,n}^t(k)$$

$$\mathbf{S}_n^t(k) = \mathbf{R} + \mathbf{H}(\hat{x}_{p,n}^t(k)) \hat{\mathbf{P}}_{p,n}^t(k) [\mathbf{H}(\hat{x}_{p,n}^t(k))]^T$$

$$\mathbf{K}_n^t(k) = \hat{\mathbf{P}}_{p,n}^t(k) [\mathbf{H}(\hat{x}_{p,n}^t(k))]^T [\mathbf{S}_n^t(k)]^{-1}$$

$$w_n^t(k) = \sum_{i=1}^{m^t(k)} \beta_i^t \mathcal{N}(z^i(k) - \phi(\hat{x}_{p,n}^t(k)), \mathbf{S}_n^t(k))$$

$$\hat{x}_n^t(k) = \hat{x}_{p,n}^t(k) + \mathbf{K}_n^t(k) \sum_{i=1}^{m^t(k)} \beta_i^t (z^i(k) - \phi(\hat{x}_{p,n}^t(k)))$$

$$\hat{\mathbf{P}}_n^t(k) = \beta_0^t(k) \hat{\mathbf{P}}_{p,n}^t(k) + (1 - \beta_0^t(k)) [\mathbf{I} - \mathbf{K}_n^t(k) \mathbf{H}(\hat{x}_{p,n}^t(k))] \hat{\mathbf{P}}_{p,n}^t(k) + \mathbf{K}_n^t(k) \left[\sum_{j=1}^{m^t(k)} \beta_j^t(k) v_i^t(k) [v_i^t(k)]^T - v^t(k) [v^t(k)]^T \right] [\mathbf{K}_n^t(k)]^T$$
 - Normalize the weights $\{w_n^t(k)\}_{n=1}^{N_p}$ (obtain $\{\hat{w}_n^t(k)\}_{n=1}^{N_p}$)
 - Using the weights $\{\hat{w}_n^t(k)\}_{n=1}^{N_p}$, resample the set $\{r_n^t(k), \hat{x}_n^t(k), \hat{P}_n^t(k)\}_{n=1}^{N_p}$
 - Obtain : $\hat{r}^t(k) = \frac{1}{N_p} \sum_{n=1}^{N_p} r_n^t(k)$ and $\hat{x}^t(k) = \frac{1}{N_p} \sum_{n=1}^{N_p} \hat{x}_n^t(k)$
4. Let $\hat{z}^t(k) = \phi(\hat{x}^t(k))$
 If $(z^t(k) - \hat{z}^t(k)) > \text{Threshold}$, $N_p = N_p + S_p$ else $N_p = N_p - S_p$
5. Go to step 2

$\beta_0^t(k)$ is the probability that all measurements in the validation gate of track t are false.

5 Simulations

Herein, the above mentioned algorithm is applied to jointly tracking and classifying two crossing targets. The targets trajectories are assumed as follows (the sampling period $\Delta = 10s$):

- Target 1: (1) The target starting from $[3000m - 10m/s \quad 10000m - 10m/s]$ at time $k = 0s$, the target runs for 25Δ with constant velocity model. (2) Executing a coordinated turn rate model with $0.15^\circ/s$ coordinated for 75Δ .
- Target 2: (1) The target starting from $[1958m - 10m/s \quad 10044m - 10m/s]$ at time $k = 0s$, the target runs for 50Δ with constant velocity model. (2) Executing a coordinated turn rate model with $-0.15^\circ/s$ coordinated for 50Δ .
- The crossing time is $t_0 = 25\Delta$.

The number of Monte Carlo runs is 50. The initial state vectors, of the MMPF, are 0.95 of the trues initial state vectors with covariance matrix $\text{diag}(100 \ 1 \ 100 \ 1)$. The initial number of particles is $N_p = 1000$ and the value of S_p in the previous algorithm is 20.

The measurement and the state covariance matrices are

$$R = \text{diag}(150^2 \ 0.005^2), \quad Q = \text{diag}(0.01^2 \ 0.01^2)$$

The mode switching probability matrix is:

$$\boldsymbol{\Pi} = \begin{bmatrix} 0.98 & 0.02 \\ 0.02 & 0.98 \end{bmatrix}$$

Moreover, one simulates both the true measurements and a set false alarm measurements. The latter are generated randomly within the probabilistic ellipsoid of gate $g = \sqrt{\gamma} = 4$ and centered in the predicted measurement for each target. The total number of false measurements is Poisson distributed with parameter λV with $\lambda = 4$.

Figure 1 illustrates an example of tracking of the targets in presence of clutter. The obtained results show that the proposed algorithm presents good performances of tracking.

The figure 2 illustrates the mode probability of the two models for each target using the CJPDA-MMPF. According to these obtained results (with only 50 independent runs), one can see in figure 2 that the models are tracked successfully for the two targets. More precisely, the changes of the mode probabilities in figure 2 are approximatively the same as the changes of the state vectors of the system. It is obvious that the algorithm works well and the classification is right and can be made easily.

The behavior of the number of particles controller is illustrated in figure 3. We note that the number of particles chosen is not enough to get the quality chosen (innovation quality) for this reason it was increased.

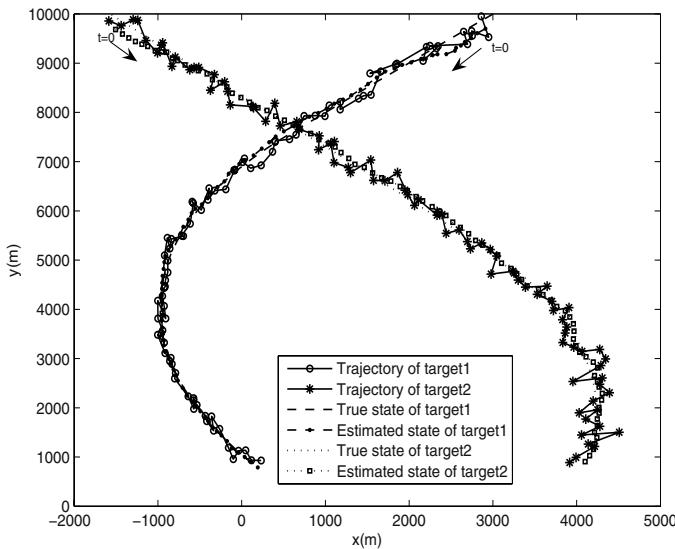


Fig. 1. Trajectories of the two targets

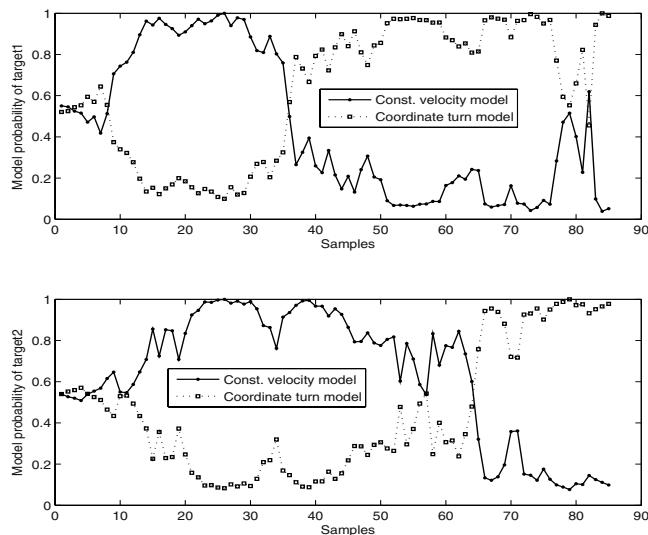
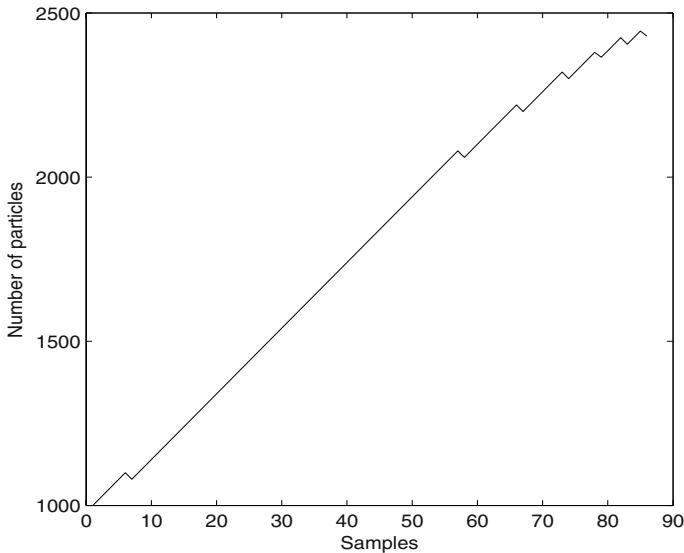


Fig. 2. The mode probability using the CJPDA-MMPF

**Fig. 3.** Number of particles

6 Conclusion

In this paper, the problem, of jointly tracking and classifying several targets in cluttered environment, is addressed using the cheap joint probabilistic data association (CJPDA) associated with the multiple model particle filter (MMPF). It is assumed that the motion model of any target belongs to one of several classes. The MMPF is used to perform nonlinear filtering with switching dynamic models and the CJDAF is used to estimate the joint probability association target / measurement. Furthermore, a controller that tunes the number of particles used by the filter using the quality of the innovation is implemented. The estimation process is therefore carried out at each particle level. The numerical simulations show interesting performances of the CJPDA-MMPF for both classification and tracking multiple targets in cluttered environment.

References

1. Blackman, S.S., Popoli, R.: Design and analysis of modern tracking systems. Artech House (1999)
2. Bar-Shalom, Y., Fortmann, T.: Tracking and Data Association. Mathematics in Science and Engineering, vol. 179. Academic Press, London (1988)
3. Bar-Shalom, Y., Xiao-Rong, Li.: Estimation and Tracking: Principles, Techniques, and Software. Artech Houes (1993)
4. Anderson, B.D.O., Moore, J.B.: Optimal Filtering. Prentice Hall, Englewood Cliffs (1979)

5. Doucet, A., De Freitas, N., Gordon, N.: Sequential Monte Carlo Methods in Practice. Springer, Heidelberg (2001)
6. Gustafsson, F.: Adaptive Filtering and Change Detection. John Wiley and Sons Ltd, Chichester (2000)
7. Ripley, B.D.: Stochastic Simulation. John Wiley, Chichester (1988)
8. Fitzgerald, R.J.: Development of practical PDA logic for multitarget tracking by microprocessor. In: Bar-Shalom, Y., Noorwood, M.A. (eds.) Multitarget-Multisensor tracking: Advanced Application, pp. 1–23. Artech House (1990)
9. Messaoudi, Z., Oussalah, M., Ouldali, A.: Data association for maneuvering targets using controlled based JPDAF particle filter. In: Proc of SMC, Irlande and United Kingdom (2007)
10. Branko, R., Sanjeev, A., Neil, G.: Beyond the Kalman filter. Particle filter for tracking application. Artech Housse, Boston (2004)

Correlation, Independance and Inverse Modeling

V. Vigneron^{1,2}

¹ IBISC-lab CNRS FRE 3190

Université d'Evry, 91020 Evry Cedex, France

vincent.vigneron@ibisc.univ-evry.fr

² Équipe MATISSE-SAMOS CES CNRS-UMR 8173

75634 Paris cedex 13,France

vigneron@univ-paris1.fr

Abstract. Learning from examples has a wide number of forms depending on what is to be learned from which available information. One of these form is $\mathbf{y} = f(\mathbf{x})$ where the input-output pair (\mathbf{x}, \mathbf{y}) is the available information and f represents the process mapping $\mathbf{x} \in \mathcal{X}$ to $\mathbf{y} \in \mathcal{Y}$. In general and for real world problems, it is not reasonable to expect having the exact representation of f . A fortiori when the dimension of \mathbf{x} is large and the number of examples is little. In this paper, we introduce a new model, capable to reduce the complexity of many *ill-posed* problems without loss of generality. The underlying Bayesian artifice is presented as an alternative to the currently used frequency approaches which does not offer a compelling criterion in the case of high dimensional problems.

1 Introduction

Inverse problems are defined, as their name points it out as the opposite of direct problems. Such definition is empty of sense as long as one do not define what is a *direct problem*. An inverse problem is a situation in which one try to determine reasons of a phenomenon from the experimental observations of his effects. For instance, in seismology, the location of the origin of an earthquake from measurements made by several seismic stations is an inverse problem. The resolution of such problem goes through an initial stage of modeling of the phenomenon, which describes the influence of the parameters of the model translated experimentally in visible effects. Then, these parameters are estimated. Mathematical resolution is made difficult due to the fact that inverse problems are in general *ill-posed* problems, *i.e* the only observations are not enough to determine perfectly all parameters of the model. It is therefore necessary to add priors which allow to reduce the space of solutions in order to lead to an unique resolution. If a problem is *well-posed*, there are good chances that a resolution is found by a stable algorithm, on a computer. Otherwise, it must be reformulated for numerical treatments. Typically, it assumes additional hypotheses, for instance the regularity of the solution. This process is known under the name of *regularization* [7].

The mathematical term of “well-posed problem” comes from a definition of Hadamard. He thought that the mathematical models of physical phenomena should have the following properties

1. one solution exists
2. the solution is single
3. resolution depends continuously on data, for a reasonable topology.

When several conclusions are consistent with the same data, the problem is to choose the best one, such as the one with minimum variance by example [3]. Although in terms of functional analysis, ill-posed problems are typically continuous, they may be unstable when they are solved with a finite precision, or with errors in the data.

2 General Inverse Model

In large scale problems (with a large number of sensors), the most common option is a reduction of dimensionality, usually performed in conjunction with a whitening step. Another approach, especially popular in the source separation community, is to work by deflation. This trail is usually followed for problems in which the number of sensors and the number of sources are both high [2]. The inherent problem with large dimension is the problem of empty space and the *curse to dimensionality*: observations quickly become very small samples in large sizes and are also noisy. The reduction is essential, but the methods are generally not robust w.r.t. noise and only applicable in a linear framework. If it is restricted to the linear framework, then one can not probably avoid taking into account the noise [4].

In source separation for instance, deflation techniques accumulate errors over time as the number of sources extracted increases [2]. The problem is important for many applications: EEG (electro-encephalography), BSPM (body surface potential mapping), MEG (magneto-encephalography), NMR (nuclear magnetic resonance), hyperspectral imaging, data analysis, etc.

Another way consists on the contrary to take advantage of the dimensionality of the data, as illustrated in Fig. II. Take the example of a non-monotone relationship $x \rightarrow y = f(x)$. In figure IIa. the solutions to the equation $y = f(x)$ at y_0 are $\{x_1, x_2, x_3\}$ and nothing can aid for distinguishing them. But, if several achievements of the relationship $y = f(x)$ are available (cf. Fig. IIb), these achievements can be combined in order to bring about a *most likely* solution than others. This example also highlights that “locality” supports the inverse approach.

When a mathematical problem is convex, it is in general quite easy to find the global minimum by using a local method. Hence the idea, in the non-convex case, is to partition the domain in order to approximate the non-convex problem on each subset of the partition, by a subset of convex problems. Most of non-convex problems can be approximated by a partition into convex problems, by “relaxing” certain constraints and the criterion [1]. A possible approach for choosing a relaxation strategy can be itemized into 4 stages :

- The choice of the *structure*: for one input vector $\mathbf{x} = (x_1, \dots, x_n)^T$ of dimension n and an output \mathbf{y} of dimension m , there is n “one-to-many” independent models $M_i, i = 1, \dots, n$, parameterized by a vector of parameters $\mathbf{p}_i \in \mathbb{R}^{\sigma(i)}$, where $\sigma(i)$ is the number of parameters for the model

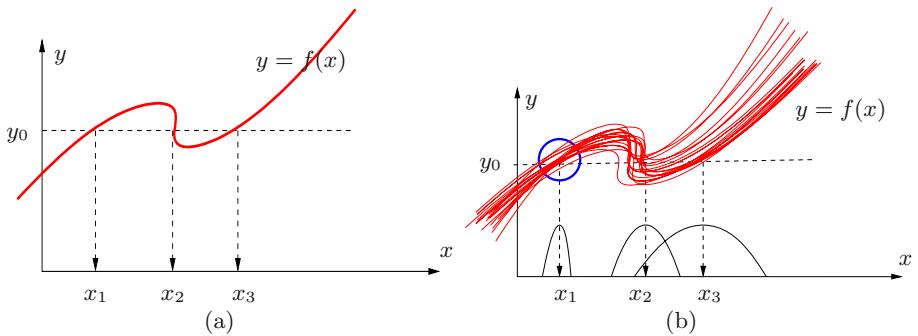


Fig. 1. (a) An instance of an ill-posed problem (b) x_0 is likely the solution sought because the density of the solutions is sharper around x_0 than around x_1 or x_3

M_i . One can have $\sigma(i) = \sigma(j), \forall i \neq j$, but not necessarily. In the following, $M(\mathbf{p}) = \{M_1(\mathbf{p}_1), M_2(\mathbf{p}_2), \dots, M_n(\mathbf{p}_n)\}$ denotes the structure of the model. Eligible prior set for the vector of parameters $\mathbf{p} = (\mathbf{p}_1, \dots, \mathbf{p}_n)^T$ of the model is \mathbb{R}^σ , where $\sigma = \sum_i \sigma(i)$. Each parameter vector \mathbf{p}_i is a model partner $M(\mathbf{p}_i)$ of the structure. The dependence between the output x_i and the inputs \mathbf{y} is linear (or not) w.r.t. the parameters \mathbf{p}_i .

- The *data collection*: vectors \mathbf{x} and \mathbf{y} are N measurements collected on a multivariable system (multiple inputs and multiple outputs). Each model $M_i(\mathbf{y}; \mathbf{p}_i)$ generates a scalar output $x_{M_i}(\mathbf{p}_i)$, homogeneous to the entry x_i . The error between the data x_i and the model output $x_{M_i}(\mathbf{p}_i)$ is defined by:

$$e(\mathbf{p}_i) \equiv e_i = x_i - x_{M_i}(\mathbf{p}_i). \quad (1)$$

- the *estimation*: it is to retain a model $M(\mathbf{p})$ generating outputs $x_{M_1}(\mathbf{p}_1), x_{M_2}(\mathbf{p}_2), \dots, x_{M_n}(\mathbf{p}_n)$ sufficiently similar to the experimental data, i.e. such that error $e(\mathbf{p})$ be small in a sense to be defined.
- the *inference*: it aims to calculate the response of the model $\hat{\mathbf{y}}$ when the vector \mathbf{x} feeds the model, combining the answers (and not vice versa). This combination of information is a *sensitive* thing that will be detailed in the Algorithm **I**. The multi-model approach allows us to exploit models $M_i(\mathbf{p}_i)$ well-suited to the local complexity of the considered sub-problems and in which a structural rule $\phi(\cdot)$ is used to get $\hat{\mathbf{y}}$. How taking advantage of the information provided by the errors \mathbf{e} to estimate $\hat{\mathbf{y}}$?

Suppose that the function $\mathbf{f} : \mathbb{R}^m \rightarrow \mathbb{R}^n$, associated to $M(\cdot)$, is continuous and locally injective. Note that the reverse function \mathbf{f}^{-1} in general is not defined (even when $n = m$). However, it is never defined everywhere. It is difficult to study the type of discontinuity that may appear around \mathbf{f}^{-1} , but we will use the following definition:

Definition 1 (stability of a solution). *The vector \mathbf{y} defines a stable point if $\partial x_{M_i}(\mathbf{y}; \mathbf{p}_i) / \partial \mathbf{y} \rightarrow 0$.*

When \mathbf{y} is unstable, there is probably a compact containing \mathbf{y} in which \mathbf{f}^{-1} is discontinuous.

3 Stochastic Search of the Solution

Starting from the previous definition, we imagined an estimation procedure based on a Markov chain structure, which can be exploited to find the solution to the problem of inference.

The estimation procedure uses a MCMC approach that emphasizes the stochastic search to resolve the problem of inference. See [5] for more details. The Markov chain structure does not contain any attributed parameters. Let $\{\partial x_{M_i(\mathbf{y}; \mathbf{p}_i)} / \partial y\}, i = 1, \dots, n$ be the jacobian values of the parcimonious models, the algorithm is the following one:

Algorithm 1. Inverse modeling using a Gibbs sampler

```

input : model:  $M_i()$ ; precision:  $\epsilon$ 
Data:  $\mathbf{x}$ ;
initialization :  $k = 0, \mathbf{y} = \mathbf{y}^{(0)}$ 
repeat
    feed-forward  $\mathbf{y}^{(k)}$ 
    for  $i=1$  to  $n$  do
        compute  $\partial x_{M_i(\mathbf{y}; \mathbf{p}_i)} / \partial \mathbf{y}, i = 1, \dots, n$ 
         $\delta \mathbf{y}^{(k+1)} = \rho(\dots, x_{i-1}^{(k+1)}, x_{i+1}^{(k)}, \dots, x_d^{(k)}, \mathbf{y}^{(k)}) \sim$ 
         $\mathcal{N}(\partial x_i^{(k)} / \partial \mathbf{y}^{(k)}(F_{\theta_i}(\mathbf{y}^{(k)}) - x_i), \sigma_{\mathbf{y}^{(k)}}^2)$ 
         $u^{(k)} \sim \mathcal{U}_{[0,1]}$ 
         $\mathbf{y}^{(k+1)} =$ 
        
$$\begin{cases} \mathbf{y}^{(k)} + \delta \mathbf{y}^{(k+1)} & \text{if } u^{(k)} < \rho(x_1^{(k+1)}, \dots, x_{i-1}^{(k+1)}, x_{i+1}^{(k)}, \dots, x_d^{(k)}, \mathbf{y}^{(k)}), \\ \mathbf{y}^{(k)} & \text{if } u^{(k)} \geq \rho(x_1^{(k)}, \dots, x_{i-1}^{(k+1)}, x_{i+1}^{(k+1)}, \dots, x_d^{(k)}, \mathbf{y}^{(k)}). \end{cases}$$

    end
     $k = k + 1$ 
until  $\|\mathbf{y}^{(k+1)} - \mathbf{y}^{(k)}\| > \epsilon$ 

```

By setting $q(\mathbf{x}^{(k)}) = \exp\left(-\sum_{i=1}^n \left|\hat{x}_{M_i(\mathbf{y}; \mathbf{p}_i)}^{(k)} - x_i^{(k)}\right|^2 / 2\sigma_x^2\right)$, Robert [5] propose to use the weighted estimation $\hat{\mathbf{y}}^{(k+1)} = \sum_k q(\mathbf{x}^{(k)}) \mathbf{y}^{(k)} / \sum_k q(\mathbf{x}^{(k)})$ where $\sum_k q(\mathbf{x}^{(k)})$ performs a normalisation of the coefficients $q(\mathbf{x}^{(k)})$. The normal law $\mathcal{N}(\partial x_i^{(k)} / \partial \mathbf{y}^{(k)}, \sigma_{\mathbf{y}}^2)$ provides a probabilistic representation of the hazard in an hypercube centered on $J_i = \partial x_{M_i(\mathbf{y}; \mathbf{p}_i)} / \partial y$ and of size $[-\sigma_{\mathbf{y}}^2, \sigma_{\mathbf{y}}^2]$. But, the random sampling stage $\delta(\mathbf{y}^{(k)})$ can be encountered more frequently under the following form:

$$\mathbf{y}^{(k+1)} = \begin{cases} \mathbf{y}^{(k)} + \delta \mathbf{y}^{(k+1)} & \text{if } u < \rho(\mathbf{p}^{(k)}, \mathbf{y}^{(k)}) \\ \mathbf{y}^{(k)} & \text{else} \end{cases}. \quad (2)$$

The sampling scheme proposed in [2] leads to no statistical bias. Robert [5] shown that the *free random walk* (constituted of these small jumps with no energetic constraint) is ergodic and therefore visits all sites in the model with the same frequency. Some remarks regarding the algorithm Tab. II: it consists to test according to an energetic eligibility criterion a series of individual displacements $\{\delta \mathbf{y}^{(k)}\}_{t>1}$ obtained by drawing for each of them [2] all involved experts, according to the value of their dynamic variables $\partial x_i / \partial y, i = 1, \dots, d$. The displacement leads, from $\mathbf{y}^{(k)}$, to $\mathbf{y}^{(k+1)}$. It can be shown that the distribution of random configuration converge to the expected distribution $q(\mathbf{x}^{(k)})$ if the number of simulations $T \rightarrow \infty$.

The Monte Carlo method makes it possible to circumvent the difficulties linked to the finite size of the model while preserving the statistical qualities of the simulation. One of the major difficulties is inevitably the bias due to the use of a discrete model to reproduce the real system.

The convergence properties in the chain $\{y^{(k)}\}$ are directly related to the Jacobien $J^{(k)}$ in the sense that $\{\mathbf{y}^{(k)}\}$ is *aperiodic* and *irreducible*.

4 Convergence Properties

Tarantola [6] and Tikhonov [7] have shown that it is not always possible to demonstrate the existence and uniqueness of the solution. To find the solution $\hat{\mathbf{y}}$ of the equation $f(\mathbf{y}) = \mathbf{x}$, let us give a candidate Lyapunov function V :

$$V = \frac{1}{2} \|f(\mathbf{y}) - \mathbf{x}\|^2, \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m.$$

Let us note J the $m \times n$ jacobian matrix defined by $\partial \mathbf{x} / \partial \mathbf{y}$ and the hessian matrix $\partial^2 V / \partial x^2$. In the following, we will assume some additional conditions on the regularity of J :

$$V \in C^2(\mathbb{R}^p) \text{ is non-constant convex,} \quad (3)$$

$$\lim_{|\mathbf{x}| \rightarrow \infty} V = +\infty \quad (4)$$

$$(\partial^2 V / \partial \mathbf{y}^2) \text{ is bounded and } \|(\partial^2 V / \partial \mathbf{y}^2)\|_\infty = M. \quad (5)$$

$$K = \sup_{\mathbf{x}} \left| \sum_{i=1}^p (\partial^2 V / \partial \mathbf{y}^2)_{ii}(\mathbf{y}) \right|. \quad (6)$$

The fact that V is non constant implies that $\|(\partial V / \partial \mathbf{y})(\mathbf{x})\| = \mathbf{0} \Leftrightarrow \mathbf{y} = \mathbf{y}^*$. The gradient of V at the point \mathbf{y} is $(\partial V / \partial \mathbf{y})$. Minimizing V gives an adaption rule which guarantees convergence towards the solution of $\hat{\mathbf{x}} - f(\mathbf{y}) = 0$ in the sense of Lyapunov. This can be written

$$\mathbf{y}^{(n+1)} = \mathbf{y}^{(n)} - \rho (\partial V / \partial \mathbf{y})^T (\mathbf{y}^{(n)}), \quad (7)$$

with ρ an adaptation factor. We can analyze in a specific way convergence properties of (7) by defining 2 other chains:

$$p_n = \exp(-V(\mathbf{y}^{(n)})) \quad \text{and} \quad \hat{\mathbf{y}}^{(n)} = \sum_{k=1}^n p_k \mathbf{y}^{(k)} / \sum_{k=1}^n p_k \quad (8)$$

The definition of (p_k) is intuitive and is designed to improve our estimation of \mathbf{y}^* . One could have chosen $p_n = \chi_{\mathcal{B}(\mathbf{x}^*, \delta)}(f(\mathbf{y}_k))$, where \mathcal{B} indicates the open ball, just to keep in the chain $\hat{\mathbf{y}}^{(n)}$ the values of \mathbf{y}_k whose images by f is distant from \mathbf{x}^* of $< \delta$.

Lemma 1. *The chain $E[V(\mathbf{y}^{(n)})]$ is decreasing and converges toward 0.*

Proof (Lemma 1). This requires the second-order Taylor-Lagrange equality when \mathbf{y} is in the vicinity of the optimum. There exists $\mathbf{z} \in [\mathbf{y}^{(n)}, \mathbf{y}^{(n+1)}]$ such that:

$$\begin{aligned} V(\mathbf{y}^{(n+1)}) - V(\mathbf{y}^{(n)}) &= \frac{\partial V}{\partial \mathbf{y}}^T (\mathbf{y}^{(n)}) \cdot (\mathbf{y}^{(n+1)} - \mathbf{y}^{(n)}) + \frac{1}{2} (\mathbf{y}^{(n+1)} - \mathbf{y}^{(n)})^T \cdot \frac{\partial^2 V}{\partial \mathbf{y}^2}(\mathbf{y}) \cdot (\mathbf{y}^{(n+1)} - \mathbf{y}^{(n)}) \\ &= -\rho (\partial V / \partial \mathbf{y})^T (\partial V / \partial \mathbf{y}) + \frac{1}{2} \rho^2 (\partial V / \partial \mathbf{y}) \cdot \frac{\partial^2 V}{\partial \mathbf{y}^2}(\mathbf{y}) \cdot (\partial V / \partial \mathbf{y})^T. \end{aligned}$$

Taking the expectation, it comes:

$$\begin{aligned} E[V(\mathbf{y}^{(n+1)}) - V(\mathbf{y}^{(n)})] &= -\rho (\partial V / \partial \mathbf{y})^2 + \\ &\quad + \frac{1}{2} \rho^2 (\partial V / \partial \mathbf{y})^T \cdot \frac{\partial^2 V}{\partial \mathbf{y}^2}(\mathbf{y}) \cdot (\partial V / \partial \mathbf{y}) + \frac{1}{2} \sum_{i=1}^p (\partial^2 V / \partial \mathbf{y}^2)_{ii} (\partial V / \partial \mathbf{y})_i^T. \end{aligned}$$

From (5) and (6), it comes:

$$E[V(\mathbf{y}^{(n+1)}) - V(\mathbf{y}^{(n)})] \leq -\rho (\partial V / \partial \mathbf{y})^2 + \frac{1}{2} \rho^2 M (\partial V / \partial \mathbf{y})^2 + \frac{1}{2} K \sigma^2. \quad (9)$$

(9) is negative iff

$$\rho \in \left[\frac{1}{M} - \frac{1}{M} \sqrt{1 - \frac{K \sigma^2}{(\partial V / \partial \mathbf{y})^2}}, \frac{1}{M} + \frac{1}{M} \sqrt{1 - \frac{K \sigma^2}{(\partial V / \partial \mathbf{y})^2}} \right].$$

In these circumstances, the chain $E(V(\mathbf{y}^{(n)}))$ is decreasing and bounded, and thus converges. \square

Lemma 2. *Let f a fonction satisfying the conditions imposed on V , and (\mathbf{x}_n) a chain in \mathbb{R}^p . For further simplifications, we suppose that $\mathbf{y}^* = 0$. Then*

$$\lim_{n \rightarrow \infty} f(\mathbf{x}_n) = 0 \Leftrightarrow \lim_{n \rightarrow \infty} \mathbf{x}_n = 0.$$

Proof (Lemma 2). The direction ' \Leftarrow ' is evident because of the continuity of f . Let us reason by absurdity on the direction ' \Rightarrow '. Suppose that \mathbf{x}_n does not go toward 0. Then $\exists \epsilon, \forall N \in \mathbb{N}, \exists n > N / \|\mathbf{x}_n\| > \epsilon$. One could extract a chain (u_n)

from (\mathbf{x}_n) such that $\forall n, \|u_n\| > \epsilon$. Consider now $m = \lim_{\mathbf{y} \in \mathbb{R}^p \setminus \mathcal{B}(0, \epsilon)} f(\mathbf{y})$. This lower limit exists as $f \geq 0$. In addition $m > 0$ because (4) means that f reaches its minimum on $\mathbb{R}^p \setminus \mathcal{B}(0, \epsilon)$ and $\mathbf{0}$ is the only point such that $f(\mathbf{0}) = 0$. There is therefore a chain $f(u_n)$ extracted from (\mathbf{x}_n) such that $\forall n, f(u_n) > m$ which is in contradiction with the hypothesis. \square

Theorem 1. *The chain $E[\mathbf{y}^{(n)}]$ converges towards \mathbf{y}^* when $n \rightarrow \infty$.*

From Jensen inequality, since V is convex $\forall n, V(E[\mathbf{y}^{(n)}]) \leq E[V(\mathbf{y}^{(n)})]$. The result is obtained by applying the lemma (2) to the chain $E[\mathbf{y}^{(n)}]$. We would like to show that the chain $E[\hat{\mathbf{y}}^{(n)}]$ converges toward \mathbf{y}^* :

$$E(\hat{\mathbf{y}}^{(n)} - \mathbf{y}^*) = \sum_{k=1}^{(n)} E \left[p_k (\mathbf{y}^{(k)} - \mathbf{y}^*) / \sum_{i=1}^{(n)} p_i \right] + \sum_{k=N+1}^n E \left[p_k (\mathbf{y}^{(k)} - \mathbf{y}^*) / \sum_{i=1}^n p_i \right] \quad (10)$$

By applying Cauchy inequality:

$$\|E[\hat{\mathbf{y}}^{(n)} - \mathbf{y}^*]\|^2 \leq \left\| \sum_{k=1}^n E \left[p_k (\mathbf{y}^{(k)} - \mathbf{y}^*) / \sum_{i=1}^n p_i \right] \right\|^2 + \sum_{k=N+1}^n \|E[(\mathbf{y}^{(k)} - \mathbf{y}^*)^2]\| \left\| E \left[\left(p_k / \sum_{i=1}^n p_i \right)^2 \right] \right\|.$$

An additional condition is necessary to demonstrate that $V(\mathbf{y}^{(k)})$ converges towards $\mathbf{0}$. A second problem is to prove that $E[(p_k / \sum_{i=1}^n p_i)^2]$ tends toward 0 , which is possible only if V is radial.

The calculus of the weighted estimator $\hat{\mathbf{y}}^{(n)} = \sum_{k=1}^n p_k \mathbf{y}^{(k)} / \sum_{k=1}^n p_k$ (see Tab. 4) is appealing. But its use can lead to disastrous results if the number of simulations is insufficient (in practice < 1000 according to Robert [5]). It is a condition for almost sure convergence $\hat{\mathbf{y}} \xrightarrow{as} E_P[\mathbf{y}]$.

It is easy to show that the convergence of the weighted estimator $\mathbf{y}^{(n)}$ is faster than the uniform estimator. Compare for instance $S_n = \sum_{k=1}^n \frac{1}{n} (\mathbf{y}^{(k)} - \mathbf{y}^*)^2$ and $\hat{S}_n = \sum_{k=1}^n p_k / \sum_{i=1}^n p_i (\mathbf{y}^{(k)} - \mathbf{y}^*)^2$. Consider the function $g(\mathbf{y}) = p(\mathbf{y}) / \sum_{i=1}^n p_i$ where $p(\mathbf{y}) = \exp(-V(\mathbf{y}))$ and $p_i = \exp(-V(\mathbf{y}_i))$. Let us denote $q_i = g(\mathbf{y}_i)$. g is defined on \mathbb{R}^p with values in \mathbb{R}_+ . Since $\forall i, p_i \leq 1$, one has $\sum_{i=1}^n p_i \leq n$, hence:

$$g(\mathbf{y}^*) = 1 / \sum_{i=1}^n p_i \geq \frac{1}{n}. \quad (11)$$

Moreover $\lim_{|\mathbf{y}| \rightarrow \infty} g(\mathbf{y}) = 0$. So $\{\frac{1}{n}\} \in g(\mathbb{R}^p)$. Let $\mathcal{V} = g^{-}([\frac{1}{n}, +\infty[)$. \mathcal{V} is bounded and contains \mathbf{y}^* . If J is radial, \mathcal{V} is a closed ball of center \mathbf{y}^* . We

denote by $\mathcal{S} = \delta\mathcal{V}$ the frontier of \mathcal{V} . As g is continuous, $\mathcal{S} = g^-(\{\frac{1}{n}\})$. If M denotes the radius of the ball \mathcal{V} , and $\mathbb{I} = \{i \in \mathbb{N}/\mathbf{y}_i \in \mathcal{V}\}$, $\mathbb{N} = \{1, \dots, n\}$, then:

$$\hat{S}_n - S_n = \sum_{i \in \mathbb{I}} \underbrace{(q_i - \frac{1}{n})}_{\geq 0} \underbrace{\|\mathbf{y}_i - \mathbf{y}^*\|^2}_{\leq M} + \sum_{i \in \mathbb{N} \setminus \mathbb{I}} \underbrace{(q_i - \frac{1}{n})}_{\leq 0} \underbrace{\|\mathbf{y}_i - \mathbf{y}^*\|^2}_{\geq M} \quad (12)$$

$$\hat{S}_n - S_n \leq M \sum_{i \in \mathbb{N}} (q_i - \frac{1}{n}) \leq 0. \quad (13)$$

A natural approach consists to examine graphically the behavior of the simulated markov chain which usually takes on a noisy aspect, see for example Fig. 2. One can plot the $\{y^{(t)}\}$ values w.r.t. t . The rapid stabilization of the estimator values indicates that convergence is reached fairly quickly: figure (2b) shows a fast convergence as soon as the first iterations.

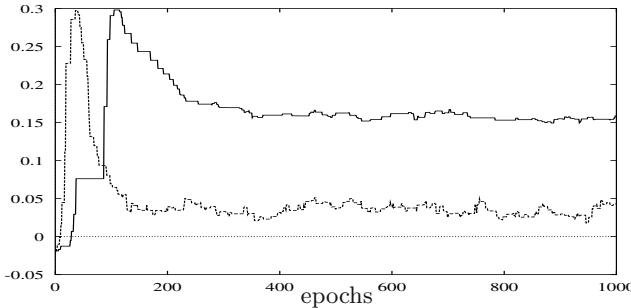


Fig. 2. (a) Markov Chain convergence of the series $\hat{\mathbf{y}}^{(k+1)} = \sum_k q(\mathbf{x}^{(k)})\mathbf{y}^{(k)} / \sum_k q(\mathbf{x}^{(k)})$ and $\frac{1}{N} \sum_{i=1}^N y^{(i)}$

5 Simulations

Example 1 (incomplete data). An interesting example of resolution concerns the (classical) case of mixtures of density. Suppose that y is written as the sum of 4-Gaussians and that we do have information only on (any) 3 of them. To what extent, our model is able to recover the fair values of y ? To study this problem, we generate a mixture of 3 Gaussians in the following manner: $x \sim \mathcal{U}_{[0,1]}$ and $y = \sum_{i=1}^4 p_i \phi_i(x)$, $\phi_i \sim \mathcal{N}(\mu_i, 1)$. At the end of the estimation stage, the numerical values of the parameters are the following : $p_i = \{0.1; 0.5; 0.3; 0.12\}$ and $\mu_i = \{0.1; 0.5; 0.7; -0.1\}$ respectively.

Now, we must find from the data triplet $(\phi_1(x), \phi_2(x), \phi_3(x))$ the value of y . At convergence, the plots Fig. 3(a-d) show that the algorithm is capable to rebuild a solution and, moreover, to estimate the sampling from $\phi_4(x)$.

Figures 3(c-d) demonstrate the differences in scale between log-linear expected/estimated inputs/outputs. Results are quite good for the prediction of \mathbf{y} , but

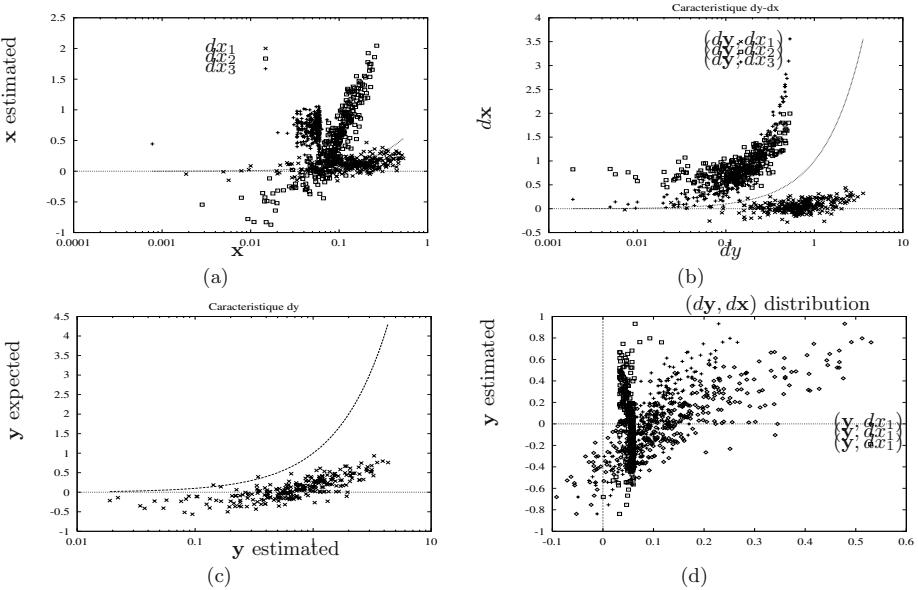


Fig. 3. Rebuild of a solution in the context of missing data. (a) plot $(x_{M_i}(\mathbf{p}_i), x_i)$ (b) plot (dy, dx) (c) plot dy (d) plot $(y, x_{M_i}(\mathbf{p}_i))$. These features are aligned with the diagonal with (few) loss of information caused by the missing variable.

poor for the entries. To counterbalance the loss of information due to the missing model ϕ_4 , the algorithm rebuild \mathbf{y} by arranging the distributions $(\phi_i)_{1 \leq i \leq 3}$. This last point is particularly easy to check on the figures 3(a-b).

6 Discussion

We presented an nonlinear approach capable to solve ill-posed problems:

- that take advantage of the dimensionality (rather than to escape)
- local in the sense that it is effective in a particular area of the signal space.

The inverse model develops an internal representation to the appropriate data structure. By cutting the space signal of its components, the algorithm avoid dimensionality constraints, but must manage as the counterpart, an inference stage that can be risky. It is a *parsimonious* model, a concept in vogue in the signal processing community. It combines heterogeneous sources, *i.e.* the information to aggregate do not necessarily come from a single random source; we do not suppose also that the sources are independent. The information provided by a source could have very different forms. In the case of aggregation of information, it is important to identify what are the reliable sources and those in error. There is no question of making the average between the two. A more natural idea is to look for the overlap between these informations in order to emphasize

areas of consensus. The more these zones are extended, the more the sources are consistent. Conversely, the fact that the sources offer conflicting information suggests that at least one of these sources provides false information. Two coinciding sources should not count for two. In addition, if a source is considered more reliable than another, the information provided by the first have priority over those in the second.

References

1. Falk, J.E.: Global solutions for signomial programs. Technical Report 6-274, George Washington Univ., Washington DC, USA (1973)
2. Hyvarinen, A., Oja, E.: Independent component analysis: algorithms and applications. *Neural Networks* 13(4-5), 411–430 (2000)
3. Mendel, J.M.: Maximum-Likelihood deconvolution. A journey into model based signal processing. Springer, New York (1990)
4. Pham, D.T., Cardoso, J.-F.: Independent Component Analysis and Blind Signal Separation (chapter Optimization issues in noisy Gaussian ICA). In: Puntonet, C.G., Prieto, A.G. (eds.) ICA 2004. LNCS, vol. 3195, pp. 41–48. Springer, Heidelberg (2004)
5. Robert, C.P.: Méthodes de simulation en statistiques. Introduction aux méthodes de Monte-Carlo par chaînes de Markov. Economica, Paris (1996)
6. Tarantola, A.: Inverse problem theory. Models for data fitting and model parameter estimation. Elsevier, Amsterdam (1987)
7. Tikhonov, A.N., Arsenin, V.Y.: Solutions of ill-posed problems. V.H. Vinsten, Washington (1977)

A New Approach for Reducing Embedding Noise in Multiple Bit Plane Steganography

Arijit Sur, Piyush Goel, and Jayanta Mukhopadhyay

Department of Computer Science and Engineering,
Indian Institute of Technology, Kharagpur
([arijits](mailto:arijits@iitkgp.ernet.in),[piyush](mailto:piyush@iitkgp.ernet.in),[jay](mailto:jay@iitkgp.ernet.in))@cse.iitkgp.ernet.in

Abstract. In this paper, a new steganographic paradigm for digital images has been proposed. In order to reduce embedding noise we propose that information should be embedded in the scaled version of a grayscale value rather than directly in the grayscale value. This approach can be combined with any multiple bit plane embedding scheme to reduce the embedding noise. We also introduce a new steganographic scheme by combining the proposed approach with an existing multiple bitplane steganographic scheme and compare the performance of the combined scheme against the bare version of the existing steganographic scheme. Experimental results reveal that for same embedding rate a multiple bit plane embedding scheme combined with proposed approach adds less embedding noise and thus is less detectable against Wavelet Moment Analysis blind steganalytic attack than its bare version.

1 Introduction

The purpose of steganography is to hide the very presence of communication by embedding messages into innocent looking cover media like digital images. Thus the security or the undetectability of a steganographic algorithm is of foremost importance. LSB replacement is a well-known steganographic scheme where the Least Significant Bits of the cover image are replaced by the message bits. No significant visual distortion can be observed in a cover image when modifications are made in the LSB planes. But LSB replacement leads to a structural imbalance in the cover image which is exploited by the steganalytic attacks [7,8] developed for the LSB replacement method.

In recent steganographic research [2,3], the LSB method has been extended to multiple bit planes (2 or 3 planes). It should be noted that even LSB embedding upto 3 bitplanes does not cause significant visual distortion. Moreover, with respect to attacks based on the structural weakness of LSB method, higher bit plane LSB embedding is more secure than single bit plane LSB embedding [4]. On the other hand multiple bitplane embedding adds more noise in the cover image than single bit plane embedding. The amount of noise added during embedding depends not only the number of changes in the cover medium but also the amount of noise added per change. In the blind steganalysis paradigm it is often considered that statistical detectability of an embedding procedure is

directly related to the total noise added during embedding. For a given medium, the steganographic algorithm which makes fewer embedding changes or adds less additive noise will be less detectable as compared to an algorithm which makes relatively more changes or adds higher additive noise [5].

The main objective of the proposed work is to suggest a new approach to data hiding which reduces the amount of embedding noise when applied jointly with some suitable multiple bit plane embedding scheme. The ability of the proposed approach to reduce noise lies in the observation that when a small number is embedded (added or subtracted) to a higher scaled version of a number(say carrier), then the embedding distortion in the carrier is less as compared to embedding the same small number into the unscaled version of the carrier. We combine this approach with an existing multiple bitplane embedding scheme and analyze the resulting noise and security of the combined scheme to validate the hypothesis.

The rest of the paper is organized as follows: in section 2, the proposed approach has been discussed, in section 3 we combine the proposed approach with an existing scheme to introduce a new steganographic scheme. In section 4, we give the detailed noise and security analysis of the combined steganographic scheme against an existing multiple bit plane scheme and finally conclusions are drawn in section 5.

2 Proposed Approach

As discussed above, the embedding distortion is less when some modification is made to a scaled version of a number and the modified number is converted back to the original scale than making the same modification to the original version of the number. When a decimal number is converted to a lower base p , ($p < 10$), then the decimal coded value of the corresponding p base number is always greater than the original decimal number. For example, consider a decimal number $x = 156$, then the equivalent base 3 number (x_3) will be 12210₃. If we consider x_3 as a decimal value, it is *twelve thousand two hundred and ten* which is clearly greater than x . So we can say that x is scaled in the p base domain. Thus any change made to the p base number is reduced when the changed p base number is converted back to decimal base. We extend this concept to steganography in order to reduce the embedding noise when multiple bit planes are used for embedding. Any embedding operation can be modeled as a mathematical function (say f). Now rather than applying f directly on a grayscale value (say x) of an image pixel for embedding, the grayvalue can be converted to a lower base number (say x_p) using Eq. 1.

$$x_p = \sum_{i=0}^{k-1} \left[\text{mod} \left(\left\lfloor \frac{x}{p^i} \right\rfloor, p \right) \right] \times 10^i \quad (1)$$

where $p < 10$ and x_p is a k digit number. x_p is considered as a decimal coded value and the embedding function f is applied to x_p producing x'_p as stego value in p base domain i.e $x'_p = f(x_p)$. Now the decimal equivalent of x'_p (say x''_p) can be computed by Eq. 2.

$$x_p'' = \sum_{i=0}^{k-1} \left[\text{mod} \left(\left\lfloor \frac{x_p'}{10^i} \right\rfloor, 10 \right) \right] \times p^i \quad (2)$$

where x_p' is a k digit number. The hypothesis is that $\text{abs}(x - x_p'') < \text{abs}(x - f(x))$ for a significant number of pixels of a grayscale image. It should be pointed out that embedding in the above fashion reduces noise for a percentage of grayscale values and not for all of the grayscale values. But it does lead to a significant reduction in the total embedding noise as explained in section 4.

In the next section we combine this approach with an existing multiple bit planes embedding scheme, 3LSB embedding, to introduce a new steganographic scheme which adds less noise and thus is a secure steganographic scheme.

3 Application Towards Steganography

In this section we describe a new steganographic scheme which is a combination of the proposed base changing approach and 3LSB embedding scheme. In this scheme, the pixel values of grayscale cover image [range 0-255] are converted to a p base system where $p < 10$. This p base number is treated as a decimal number and is modified to embed 3 message bits. After embedding if the modified pixel value remains a valid p base number and the noise added in case of p base embedding is less than normal *decimal base* embedding then embedding is allowed in p base system and the stego p base number is converted back to the corresponding *decimal* base. Otherwise embedding is done in the normal *decimal* base system. A shared secret parity function is used to convey the information whether embedding is in p base or *decimal* base. For our experiments we use the *XOR* operation as the parity function. We compare the proposed scheme with bare 3LSB scheme in order to determine the effectiveness of the proposed approach.

3.1 Embedding Algorithm

The algorithm for embedding l message bits in a group of pixels is outlined below. The base p is known apriori to both the encoder and the decoder.

- Divide the cover image into non overlapping groups of 2 pixels each. These groups can be formed using a pseudo-random sequence with a shared secret key.
- Consider l message bits and compute their decimal equivalent s . The first pixel of the group (let it be denoted by x) is converted to a p base system using Eq. 1 (let the p base number be x_p). x_p is treated as decimal coded number.
- The message symbol s ($0 \leq s < m$ where $m = 2^l$) is embedded in x_p and also in x using Eq. 3 which is the standard embedding function for l bitplane LSB replacement method. Let the resulting values of x_p and x be x'_p and x' respectively.

$$x' = (\left\lfloor \frac{x}{m} \right\rfloor \times m) + s \quad (3)$$

- where x' is the pixel value after embedding, $m = 2^l$ and s is the decimal equivalent of the l message bits. In our case $l = 3$ and $m = 8$.
- x'_p may not be a valid p base number or noise added due to p base system may be greater than the noise added due to the decimal system. In such a case, the group is unsuitable for p base embedding and message is embedded using normal decimal base i.e x' is considered as the stego value.
 - After embedding, if the resultant stego value (x_p) is a valid p base number then its decimal equivalent (x''_p) is calculated using Eq. 2 and is stored as stego pixel.
 - Modify the LSB of second pixel of the group to embed the parity information using the XOR function i.e. $XOR(lsb(x_1), lsb(x_2))$ is equal to 1 and 0 for p base embedding and decimal embedding respectively where x_1 and x_2 are the pixels of the group.

The extraction of the message sequence is a simple inverse process of the above algorithm as explained in the following example.

3.2 Illustrative Example

Consider two pixels (a_1, a_2) having values 169 and 170 respectively and we want to embed the message sequence 110 (decimal value 6) using a base 4 system. a_1 is first changed to its base 4 representation, which is 2221, using Eq. 1. The message sequence is embedded in 2221 using Eq. 3 which gives 2222 as the embedded number. The corresponding base 10 equivalent of this number is 170. Thus a noise of 1 is added which is much less as compared to noise added when embedding is done directly in 169 (a noise of 5 is added when embedding is done in base 10). Also we change a_2 to 171 to convey the parity information since $XOR(lsb(170), lsb(171)) = 1$. On the decoding end, we first check the parity information to find out whether decimal base embedding or p – base embedding has been done. We see that $XOR(lsb(170), lsb(171)) = 1$ hence embedding has been done in p – base. So convert 170 to its base 4 representation i.e 2222 using Eq.1 and extract the decimal equivalent of the message sequence as 2222 modulo 8 (which equals 6). It can be checked that if value of a_1 and the message sequence was 164 and 101 respectively, then base 4 embedding would have added more noise as compared to decimal base embedding and hence embedding would have been done in decimal base system.

4 Noise Analysis and Experimental Results

4.1 Noise Analysis

The total noise due to above scheme comprises of the noise due to embedding in the p base domain, the noise due to embedding in the decimal domain and

the noise due to the parity information. The reduction in noise depends on the success rate of $p - \text{base}$ embedding. Since we don't have an exact statistical model which accurately relates the success rate of p base embedding and the reduction in noise, we provide empirical results observed from experiments on one hundred test images. On testing the algorithm on one hundred test images it was observed that on average, embedding in p base system, for varying values of p , was successfully accomplished for a significant number of pixels as shown in Table 1. The total noise is reduced in the new scheme as compared with bare 3LSB scheme as reflected in the scatter plot shown in Fig 1. x axis and the y axis depict the noise per pixel due to embedding by 3LSB and the proposed scheme respectively. Scatter plot was generated using $p = 4$.

Table 1. Success Rate of $p - \text{base}$ embedding for 100 images for different bases

p base	3	4	5	5	7	8	9
Success Rate	24.45%	22.99%	29.23%	20.77%	37.72%	30.60%	39.62%

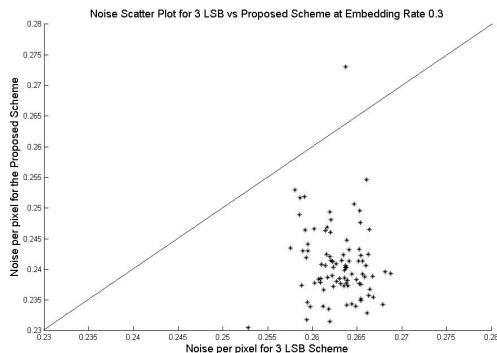


Fig. 1. Noise Scatter Plot Proposed Scheme vs 3LSB at Embedding Rate 0.30 bits per pixel. x axis:Noise/pixel for 3LSB, y axis:Noise/pixel for Proposed scheme.

We also compare the performance of the proposed scheme against 3LSB with respect to the embedding efficiency which is defined as the expected number of embedded message bits per one embedding change [11]. Table 2 gives a comparative analysis of 3LSB and proposed scheme for 2 values of pixel base p ($p = 4$ and $p = 8$) with varying embedding rates.

4.2 Security Analysis

A steganographic scheme is considered secure if given the full knowledge of the embedding algorithm, including the statistical properties of the source of the cover image, except the stego key, there are no artifacts in the stego image that

Table 2. Average Additive Noise and Embedding Efficiency for 3LSB and proposed scheme

Embedding	3LSB		Proposed Scheme Base = 4		Proposed Scheme Base = 8	
	Rate	Average Noise	Embedding Efficiency	Average Noise	Embedding Efficiency	Average Noise
0.10	0.0874	1.1456	0.0800	1.2535	0.0736	1.3565
0.20	0.1757	1.1386	0.1590	1.2541	0.1476	1.3578
0.30	0.2630	1.1411	0.2386	1.2593	0.2213	1.3581
0.40	0.3504	1.1417	0.3176	1.2609	0.2948	1.3584
0.50	0.4376	1.1427	0.3964	1.2625	0.3684	1.3590

can be detected by an attacker with a probability better than random guessing (Kerckhoffs principle) [1]. For evaluating the security of the proposed scheme, Wavelet Absolute Moment Steganalysis (WAM)[6] has been used as the steganalyzer. WAM is a blind steganalyser which classifies images as cover or stego based on a multi-dimensional feature space. The features for steganalysis are extracted from the noise component of the stego image in the wavelet domain. Assuming that the noise due to embedding is a stationary gaussian signal with a known variance, Wiener filter is used to extract noise component in the wavelet domain. All the features (statistical moments) are calculated as higher order moments of the noise residual in the wavelet domain. The total number of features for a grayscale image is $3n_{mom}$ where $n_{mom} = 9$. For classification, the 27 dimensional feature space is reduced to single dimension using Fisher Linear Discriminant (FLD) analysis. Then Linear Discriminant Analysis (LDA) classifier [10] is used to classify the projected points on the principal component axis and Receiver Operating Characteristic curves are plotted for evaluating the performance of the classifier on a steganographic algorithm.

4.3 Experimental Results

Two thousand uncompressed grayscale TIFF images of different sizes [9] were used for our experiments. The stego images were generated using the proposed steganographic scheme for base 4 (i.e $p = 4$) and bare 3LSB embedding scheme for embedding rates of 0.1 and 0.3 bits/pixel. WAM classifier was trained with 1000 cover and 1000 stego images. ROC plots were generated for comparing the combined steganographic scheme with bare 3LSB embedding. It can be observed from the ROC plots shown in Fig. 2 and Fig. 3 that against WAM based steganalyzer the 3LSB scheme combined with proposed approach is less detectable than bare 3LSB embedding scheme. The proposed scheme generates more number of false positives than 3LSB and thus the classification procedure is closer to random guessing for proposed scheme. It should be mentioned that for higher embedding rates 3LSB scheme is detected fully by WAM due to large additive noise and hence the ROC plots for higher embedding rates have not been considered.

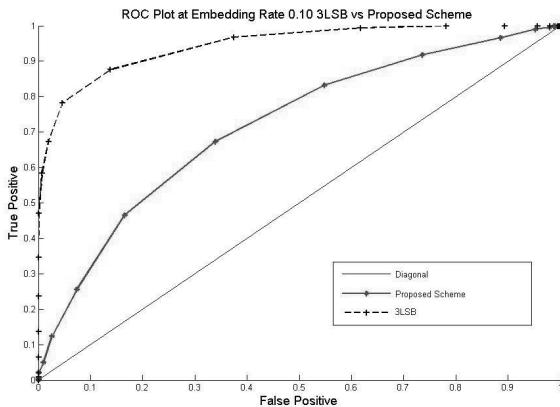


Fig. 2. ROC Plot:Proposed Scheme vs 3LSB Replacement at Embedding Rate 0.10 bits/pixel

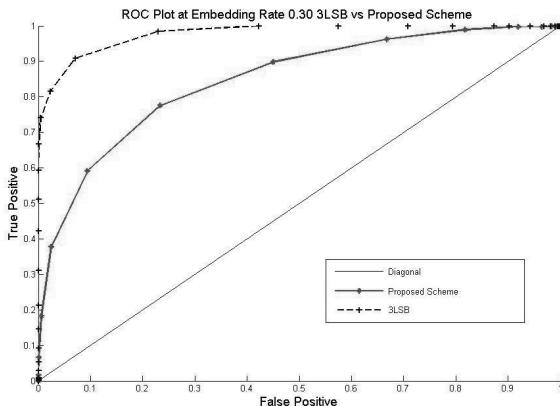


Fig. 3. ROC Plot:Proposed Scheme vs 3LSB Replacement at Embedding Rate 0.30 bits/pixel

5 Conclusion

In this paper, a new approach has been proposed which reduces the embedding noise when combined with a suitable multiple bit plane steganographic scheme. It has been shown empirically that when combined with 3LSB scheme the proposed method helps in reducing the embedding noise than bare 3LSB scheme and thus the merged scheme is statistically less detectable against WAM based steganalytic attack. Hence, the contribution of this paper is two-fold. Firstly a generalized base change method has been proposed which can be combined with different steganographic schemes. Secondly a secure steganographic

algorithm has been introduced by combining the proposed approach with an existing steganographic scheme.

References

1. Katzenbeisser, S., Petitcolas, F.A.P.: Defining Security in Steganographic Systems. In: SPIE Security and Watermarking of Multimedia Contents IV. Electronic Imaging 2000, San Jose, CA, vol. 4675, pp. 50–56 (2002)
2. Zhang, X., Wang, S.: Steganography using multiple-base notational system and human vision sensitivity. *Signal Processing Letters* 12(1), 67–70 (2005)
3. Wu, D.-C., Tsai, W.-H.I.: A Steganographic method for images by pixel-value differencing. *Pattern Recognit. Lett.* 24, 1613–1626 (2003)
4. Ker, A.: Steganalysis of Embedding in Two Least-Significant Bits. *IEEE Transaction on Information Forensics and Security* 2(1) (March 2007)
5. Fridrich, J., Soukal, D.: Matrix Embedding for Large Payloads. In: Proc. of SPIE Electronic Imaging, Photonics West (January 2006)
6. Goljan, M., Fridrich, J., Holotyak, T.: New Blind Steganalysis and its Implications. In: Proc. SPIE Electronic Imaging, Photonics West (January 2006)
7. Dumitrescu, S., Wu, X., Wang, Z.: Detection of LSB steganography via sample pair analysis. In: Petitcolas, F.A.P. (ed.) IH 2002. LNCS, vol. 2578, pp. 355–372. Springer, Heidelberg (2003)
8. Fridrich, J., Goljan, M., Du, R.: Reliable detection of LSB steganography in color and grayscale images. In: Proc. ACM Workshop on Multimedia and Security, pp. 27–30 (2001)
9. Wang, J., Li, J., Wiederhold, G.: SIMPLICITY: Semantics-sensitive integrated matching for picture LIBraries. *IEEE Trans. Pattern Anal. Mach. Intell.* 23(9), 947–963 (2001)
10. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, San Diego, California (1990)
11. Fridrich, J., Lisonek, P., Soukal, D.: On Steganographic Embedding Efficiency. In: 8th International Workshop Information Hiding, Alexandria, VA

Region-Based Wavelet-Packet Adaptive Algorithm for Sparse Response Identification

Odair A. Noskoski and José C.M. Bermudez

Federal University of Santa Catarina

Department of Electrical Engineering

odair@cefetrs.tche.br, j.bermudez@ieee.org

Abstract. This paper proposes a new wavelet-packet-based algorithm for sparse response identification. The distinctive features of the new algorithm are a region-based strategy for adaptive weight activation and a new deactivation/activation schedule across the transform scales. The new algorithm shows improved performance when compared to existing wavelet-based algorithms with similar characteristics. The new strategies lead to a faster wavelet-packet transform construction and to improved robustness to design parameters when compared to previous solutions. Monte Carlo simulation results show good performances regarding convergence speed and robustness to design parameter choice.

Keywords: Adaptive systems, echo cancellation, sparse systems.

1 Introduction

Identification of systems with sparse impulse responses is useful in wireless communications, echo cancellation, underwater acoustics and geophysics. Several algorithms exploit sparsity to improve identification efficiency by detecting the significant (active) samples of the unknown response [1, 2, 3, 4, 5, 6].

The solution in [2] is able to identify responses with an unknown number of dispersive regions. It requires a number of preset parameters, which tends to render optimal designs difficult. One of these parameters is the maximum number of active taps. Thus, efficiency is compromised unless a good estimate of this number is available. Ref. [3] incorporates a least-squares (LS) based active tap detection to the NLMS algorithm. Computational complexity is a little above NLMS's if two previously evaluated look-up tables are used. It has been reported that the LS-based detection may fail for impulse responses with large dynamic ranges [1]. Ref. [4] was the first to exploit the wavelet transform (WT) time hierarchy in sparse system identification. The Haar-Basis (HB) algorithm [4] works from a control scale for which all weights are adapted at every adaptation interval (AI). After each AI, the converged control weights which are larger than a detection threshold activate the weights in the same time hierarchy at the other scales. Then, a new AI begins during which active and control weights are adapted. This approach can be used to identify sparse responses with more than one dispersive region. Because the weight activations are based on a fixed control

scale, the algorithm may fail for sparse responses which are not rich enough in frequency content [4] or are orthogonal to the control scale basis vectors [4]. Examples can be found among typical network echo path responses [4,6]. Finally, the necessary adaptation of all the control weights may lead to longer than necessary adaptive filters, increasing computational cost and convergence time. In [5] two short adaptive filters are used. A first filter operates in a partial-Haar transform domain and estimates the location of the peak of the unknown response. The second filter is a short time domain adaptive filter centered about the estimated peak location. Positioning of the second filter uses the time hierarchy of the WT. The approach requires an estimate of the number of active coefficients and is effective for sparse responses with only one dispersive region.

Wavelet packets generalize the wavelet theory to allow flexible frequency domain signal representations [8]. Ref. [6] proposed a wavelet-packet-based (WPB) algorithm for identification of sparse impulse responses with arbitrary frequency spectra. The WPB algorithm uses localization information in both time and frequency. Frequency information is used to adaptively design a discrete wavelet packet transform (DWPT) tailored to the spectral energy distribution of the unknown response. Time information is used to adaptively locate the active weights. WPB was shown to match the results of [4] whenever HB provided good performance. Moreover, WPB provided very good results for responses not suitable for the WT-based approach of [4].

This paper proposes a new wavelet-packet-based algorithm with improved performance when compared to [4,6]. The distinctive features of the new algorithm are based on the following considerations: 1) Most practical sparse responses are composed of short and dense regions of active samples, 2) Wavelet domain representations of typical impulse responses may have different forms for different delays of the effective response [8, Page 147], 3) Long adaptive intervals, which are desirable for accurate active weight estimations, can be combined with frequent weight deactivations to speed-up convergence and reduce complexity. The new algorithm is called Region-Based Wavelet-Packet (RBWP) algorithm.

Unlike HB and WPB, which activate individual weights, RBWP activates regions of weights. This new active region (as opposed to active weight) detection strategy allows for a significant reduction of the required adaptation intervals. RBWP can build the WP transform in about 1/2 of the iterations required by WPB. RBWP uses also a more effective weight deactivation/activation schedule, as compared to WPB. The new schedule allows for a better estimate of the active weights before testing for deactivation. Monte Carlo (MC) simulations show improved convergence and robustness to design parameter choice of RBWP relative to HB and WPB. RBWP has been successfully tested with all echo responses in [7] using different effective response delays.

The paper is organized as follows. Section 2 briefly reviews the WPB algorithm [6]. Section 3 presents the new region-based activation strategy and weight deactivation/activation schedule that differentiate RBWP from WPB. The computational complexities of RBWP, NLMS, HB and WPB are compared. Section 4 presents MC simulation results and Section 5 concludes the paper.

2 WPB Algorithm

WPB has two phases [6]. In Phase 1 the DWPT is built and a first pass is made in estimating the active weights. Phase 2 refines the detection and estimation of the active weights for the DWPT defined in Phase 1.

2.1 Phase 1: Construction of the DWPT

Fig. 1 shows an M -level DWPT construction for $M = 3$. $\tilde{y}(n)$ is the estimate of the desired signal $y(n)$, which is the unknown system output plus an i.i.d. additive noise with power σ^2 . $e(n)$ is the estimation error. Phase 1 starts with a one-level DWPT realized by Haar filters \mathbf{H}_L (lowpass) and \mathbf{H}_H (highpass). Input vector $\mathbf{x}(n)$ is $N \times 1$. Decimated output vectors \mathbf{z}_{H_1} and \mathbf{z}_{L_1} are $N/2 \times 1$.

Next step is the first AI, AI₁. Two $N/2 \times 1$ adaptive filters with weight vectors \mathbf{w}_{H_1} and \mathbf{w}_{L_1} are adapted, where the subscript 1 indexes the DWPT level. The converged weights are then compared to a threshold. Those larger than the threshold are considered to be active weights. This concludes AI₁. The second AI (AI₂) follows in which only the active weights are adapted. These weights form vectors $\tilde{\mathbf{w}}_{H_1}$ and $\tilde{\mathbf{w}}_{L_1}$, which are excited by the pruned transformed signal vectors $\tilde{\mathbf{z}}_{H_1}$ and $\tilde{\mathbf{z}}_{L_1}$. With much less weights than in AI₁, AI₂ has reduced complexity and convergence time.

Next step of Phase 1 generates the next level of the DWPT. Only the subband with higher energy is again subdivided. Subband energies are estimated by $\|\tilde{\mathbf{w}}_{H_1}\|_2$ and $\|\tilde{\mathbf{w}}_{L_1}\|_2$. The weights in the lower energy subband are migrated into a new adaptive vector $\tilde{\mathbf{w}}_a$, which is excited by signal vector $\tilde{\mathbf{z}}_a$ with the corresponding signal samples. Weight vector $\tilde{\mathbf{w}}_a$ will always be adapted from this point on. The higher energy subband is subdivided by a new set of filters \mathbf{H}_L and \mathbf{H}_H . This leads to transformed signal vectors $\tilde{\mathbf{z}}_{H_2}$ and $\tilde{\mathbf{z}}_{L_2}$ at the output of the second DWPT level. The corresponding weight vectors $\tilde{\mathbf{w}}_{H_2}$ and $\tilde{\mathbf{w}}_{L_2}$ are initialized by filtering the higher energy subband coefficients ($\tilde{\mathbf{w}}_{H_1}$ or $\tilde{\mathbf{w}}_{L_1}$) with \mathbf{H}_L and \mathbf{H}_H . A new AI follows and another decision is made about which subband will be split. The process is repeated for a new DWPT level until the complete DWPT is obtained.

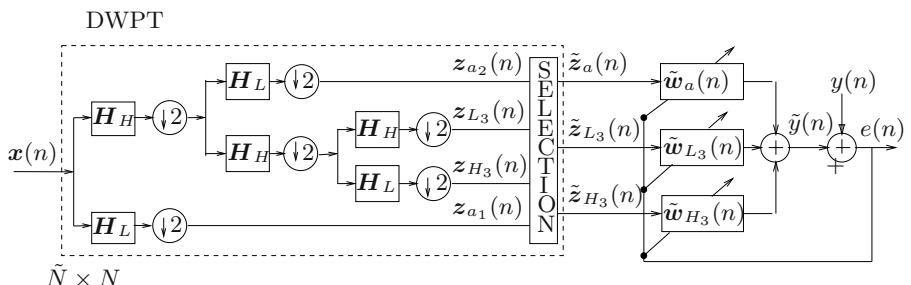


Fig. 1. Block diagram with $M=3$ levels

Any adaptive algorithm can be used. Here we use the NLMS algorithm [4]. In Fig. 1 $\tilde{y}(n) = \tilde{\mathbf{z}}^T(n)\tilde{\mathbf{w}}(n)$, $\tilde{\mathbf{z}}(n) = [\tilde{\mathbf{z}}_a^T(n), \tilde{\mathbf{z}}_{L_m}^T(n), \tilde{\mathbf{z}}_{H_m}^T(n)]^T$, $\tilde{\mathbf{w}}(n) = [\tilde{\mathbf{w}}_a^T(n), \tilde{\mathbf{w}}_{L_m}^T(n), \tilde{\mathbf{w}}_{H_m}^T(n)]^T$ and $e(n) = y(n) - \tilde{y}(n)$. The update equations are:

$$\begin{aligned}\tilde{\mathbf{w}}_a(n+1) &= \tilde{\mathbf{w}}_a(n) + 2\tilde{\mu}\Lambda_a^{-2}(n)e(n)\tilde{\mathbf{z}}_a(n) \\ \tilde{\mathbf{w}}_{L_m}(n+1) &= \tilde{\mathbf{w}}_{L_m}(n) + 2\tilde{\mu}\lambda_{L_m}^{-2}(n)e(n)\tilde{\mathbf{z}}_{L_m}(n) \\ \tilde{\mathbf{w}}_{H_m}(n+1) &= \tilde{\mathbf{w}}_{H_m}(n) + 2\tilde{\mu}\lambda_{H_m}^{-2}(n)e(n)\tilde{\mathbf{z}}_{H_m}(n)\end{aligned}$$

where $\Lambda_a^2(n)$ is a diagonal matrix of estimates of the power of $\tilde{\mathbf{z}}_a$ given by $\lambda_{a_m}^2(n) = (1-\alpha)\lambda_{a_m}^2(n-1) + \alpha\tilde{z}_{a_{m,1}}^2(n)$, $0 < \alpha < 1$, where $m = 1, 2, \dots, (M'+1)$ with $M' < M$ being the level corresponding to the present adaptive interval. $\tilde{z}_{a_{m,1}}(n)$ is the first active element of the transformed input vector for level m .

2.2 Phase 2: Adaptation

The adaptive weights defined by the DWPT constructed in Phase 1 are iteratively adapted and activated/deactivated in Phase 2 according to the temporal hierarchy of the DWPT. Fig. 2 illustrates this hierarchy for $M = 5$. The dark rectangles belong to the same temporal hierarchy of element (3,2) (marked with *). Horizontal direction shows distribution in time. Vertical direction shows the 5 levels ($m = 1, \dots, 5$). Each rectangle corresponds to the region of greater influence of a transformed weight.

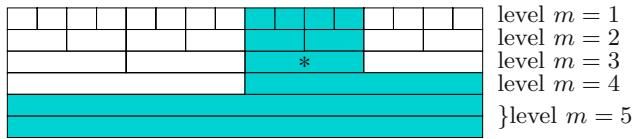


Fig. 2. Hierarchic structure of the DWPT-Haar

WPB processes the input \mathbf{x} through the DWPT to determine the reduced transformed vector $\tilde{\mathbf{z}}_a$ that will excite the adaptive filter with weight vector $\tilde{\mathbf{w}}_a(n)$, which contains the active weights of all levels. Now, $\tilde{y}(n) = \tilde{\mathbf{z}}_a^T(n)\tilde{\mathbf{w}}_a(n)$ and $e(n) = y(n) - \tilde{y}(n)$. The weight update equation is

$$\tilde{\mathbf{w}}_a(n+1) = \tilde{\mathbf{w}}_a(n) + 2\tilde{\mu}\Lambda_a^{-2}(n)e(n)\tilde{\mathbf{z}}_a(n).$$

3 RBWP Algorithm

This section describes the differences between the WPB and the proposed RBWP algorithm.

3.1 New Region-Based Activation Strategy

HB or WPB detect active weights. RBWP detects active regions. After large active weights are detected, neighboring weights about the detected weights are

aggregated to form active regions. Fig. 3 illustrates the strategy. A sub-region R_i at level m is defined about each detected weight. The union R of all R_i defines the region of weights to be declared active at level m . Weights outside R are deactivated. Weights at level $m - 1$ which are in the same time hierarchy of the active region of level m are activated. The spread of each R_i about a detected weight is a design parameter. Here we consider one unit to each side, as shown in Fig. 3. This new strategy leads to a significant reduction in the

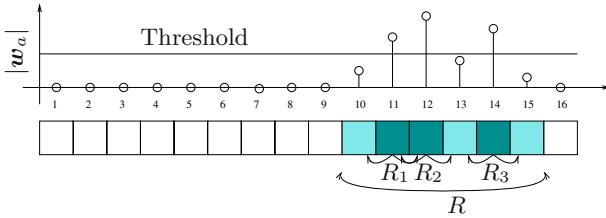


Fig. 3. Example of detection regions for a real response

number of iterations required to build the DWPT in Phase 1. RBWP requires only about 200 iterations for AI₁, the most critical step in building the DWPT, for all responses in [7]. This is an order of magnitude faster than WPB. AI₁ is then followed by several AIs using the same strategy. After some AIs it is even possible to conclude the DWPT design without new AIs. Only a succession of subband energy estimations and higher energy subband splittings is required, all based on the active samples already identified. The new strategy typically reduces the number of iterations required for WPB to build the DWPT by about 50%. Thus, the new region-based activation strategy completely modifies Phase 1, as compared to WPB.

3.2 New Weight Deactivation/Activation Schedule

Each AI in Phase 2 of RBWP is followed by a region-based weight deactivation/activation step. Besides being region-based, a new schedule is proposed which allows for larger active weight estimation time. Table II shows the WPB and the new schedule for $M = 5$ and for each cycle of AIs. Take, for instance, level 2 for the RBWP algorithm. The weight activation for this level is done after AI₃. Then, the estimation of the weights in this level will be improved over 4 AIs before they are tested for deactivation in AI₂ of the next adaptation cycle. In comparison, weights at $m = 2$ for the WPB algorithm are activated after AI₄ and tested for deactivation right after the next AI (AI₅). In HB [4] weights are also tested for deactivation after each AI. The new schedule clearly allows for either shorter AIs or improved active weight estimations before deactivation tests. This permits a better tradeoff between speed and robustness to design in a given implementation, as compared to HB or WPB.

Table 1. Deactivation/activation schedule after each AI in one adaptation cycle of WPB and RBWP algorithms for M=5

Adaptation interval in each cycle		1	2	3	4	5
WPB	Deactivation in level	1	5	4	3	2
	Activation in level	5	4	3	2	1
RBWP	Determination of R in level	1	2	3	4	5
	Activation in level	5	1	2	3	4

3.3 Computational Complexity

The exact computational complexity of HB, WPB or RBWP is very difficult to predict. The number of adapted weights varies according to decisions made during identification. Computations required for detections, decisions and step size updatings are implementation dependent. For an N -tap response ($N = 2^M$), the number \tilde{N} of weights effectively adapted (NWEA) is, in general, different for each algorithm.

Table 2 shows the number of operations required for a typical iteration of the NLMS [9], HB, WPB and RBWP algorithms as functions of N and \tilde{N} . These expressions will be used in the next section for comparison.

Table 2. Computational complexity of the NLMS, HB, WPB, and RBWP algorithms

Algorithm	Additions	Multiplications	Divisions
NLMS	$2N + 3$	$2N + 3$	1
HB-WPB-RBWP	$2\tilde{N} + 3M + 1$	$2\tilde{N} + 5M + 4$	$M + 1$

4 Simulation Results

This section compares the performances of RBWP, HB and WPB for response gm_1 in [7] using $N = 512$ taps. This is the response from [7] that led to the best results for HB. The detection threshold used for all algorithms was $TH = \beta_{fa}\sqrt{\tilde{\mu}\tilde{\xi}(k)/\tilde{\lambda}_{a_m}^2(k)}$ [4], where $\tilde{\xi}(k) = (1 - \tilde{\mu})\tilde{\xi}(k - 1) + \tilde{\mu}e^2(k)$ is an estimate of $E[e^2(k)]$. $\tilde{\mu}/\tilde{\lambda}_{a_m}^2(k)$ is the NLMS step size and β_{fa} determines the probability of false alarm [4, Section II-C]. For best performance, $\beta_{fa} = 2.56$ for HB and RBWP, and $\beta_{fa} = 0.9$ for WPB. $\mathbf{x}(n)$ was generated by passing a white, Gaussian random sequence through the filter $H(z) = 0.25\sqrt{3}/(1 - 1.5z^{-1} + z^{-2} - 0.25z^{-3})$ to produce a signal with a spectrum closely resembling the spectrum of speech [4, Eq. (29)]. The SNR (σ_x^2/σ^2) was 40dB.

Step sizes were $\tilde{\mu} = 1/10\tilde{N}$ for WPB and RBWP and $\tilde{\mu} = 1/8\tilde{N}$ for HB. AIs with $4\tilde{N}$ iterations were used for HB, RBWP and for Phase 2 of WPB. For DWPT building in WPB, AI₁ and AI₂ of Phase 1 had 2000 and 4000 iterations, respectively. Phase 1 of RBWP AI₁ had 200 iterations. All remaining AIs had

$4\tilde{N}$ iterations, up to a total of 3000, which is 50% of the number of iterations required for Phase 1 of WPB.

Figs. 4(a) and 4(c) show model gm_1 of [7] with delays $\delta = 253$ and 246, respectively, for a 512-tap channel. Figs. 4(b) and 4(d) show the corresponding WT weights for level 3, the control scale used for HB. Note that the transformed responses are quite distinct [8, Page 147]. Fig. 4(e) clearly shows that the weight at position 33 will be considered inactive for this control scale. The WT time hierarchy will then render weights 129-132 of level 1 also inactive using the HB algorithm. Fig. 4(f) shows the weights for level 1 of the WT corresponding to Fig. 4(c). It is clear that weights 129-131 should be active. Thus, a poor performance is expected for HB in this case. A similar situation occurs for WPB for $\delta = 253$ due to the quasi-orthogonality of gm_1 and the DWPT used. Both drawbacks are avoided in RBWP because of the active region detection strategy.

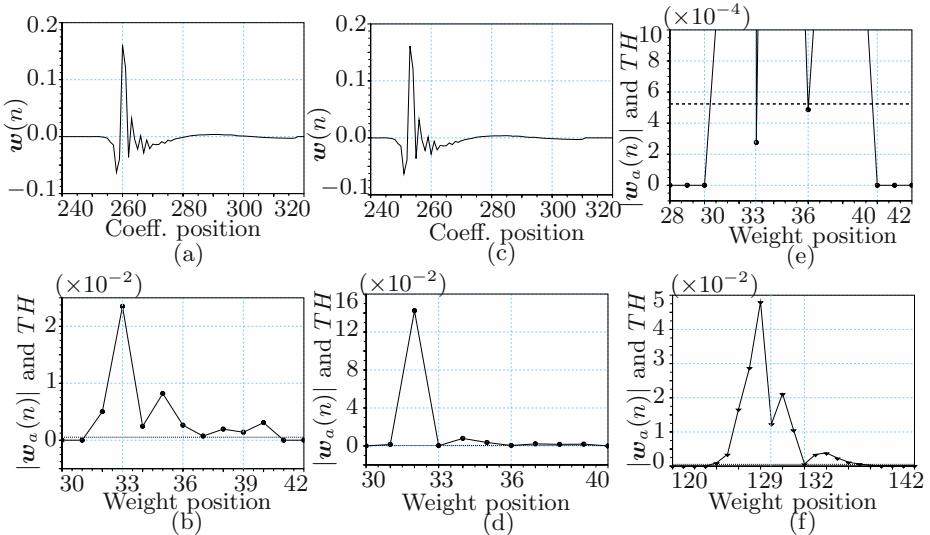


Fig. 4. (a) Model gm_1 with delay of $\delta = 253$. (b) WT weights for level 3 of signal in (a). (c) Model gm_1 with delay of $\delta = 246$. (d) WT weights for the level 3 of signal in (c). (e) ZOOM of (d). (f) WT weights for level 1 of signal in (c).

Fig. 5(a) and 5(b) show MC simulations (100 runs) for the excess mean-square error (EMSE) for $\delta = 253$ and $\delta = 246$, respectively, for NLMS (time-domain), NLMS-TD (WT-domain), HB, WPB and RBWP algorithms. Note that RBWP presents faster convergence and lower EMSE in both cases. Except for NLMS and NLMS-TD, HB (WPB) had the poorest performance for $\delta = 246$ ($\delta = 253$). These results were expected as explained above.

The average NWEA (steady-state) for both delay values were $\tilde{N} = 140$ for HB, $\tilde{N} = 60$ for WPB, $\tilde{N} = 90$ for RBWP. Using Table 2, it is verified that HB, WPB and RBWP require about 31%, 16% and 22%, respectively, of the complexity required by time-domain NLMS.

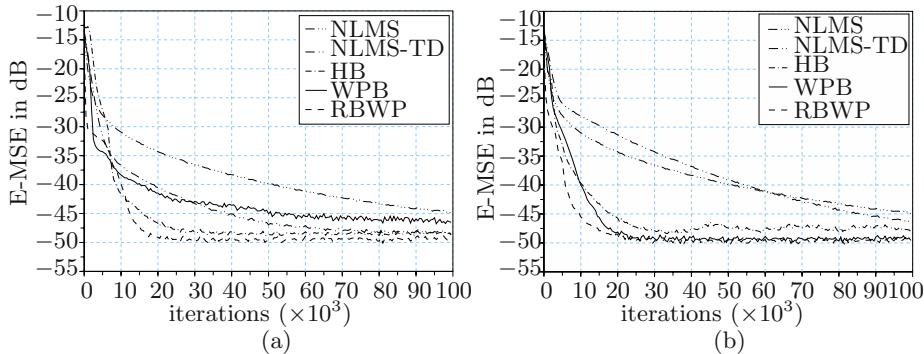


Fig. 5. Excess mean square error. (a) $\delta = 253$. (b) $\delta = 246$.

5 Conclusion

This paper has presented a new region-based wavelet-packet (RBWP) algorithm for identification of sparse impulse responses with arbitrary spectra. RBWP new region-based activation strategy significantly reduces the time required to build the wavelet-packet transform. MC simulations have shown better performance and robustness to design parameters when compared to existing algorithms.

References

1. Deng, H., Doroslovački, M.: Proportionate adaptive algorithms for network echo cancellation. *IEEE Transactions on Signal Processing* 54(5), 1794–1803 (2006)
2. Sugiyama, A., Ikeda, S., Hirano, A.: A fast convergence algorithm for sparse-tap adaptive FIR filters identifying an unknown number of dispersive regions. *IEEE Transactions on Signal Processing* 50(12), 3008–3017 (2002)
3. Homer, J., Mareels, I.: LS detection guided NLMS estimation of sparse systems. In: Proc. of the ICASSP., vol. 2, pp. 861–864 (2004)
4. Ho, K.C., Blunt, S.D.: Rapid identification of a sparse impulse response using an adaptive algorithm in the Haar domain. *IEEE Transactions on Signal Processing* 51(3), 628–638 (2003)
5. Bershad, N.J., Bist, A.: Fast coupled adaptation for sparse impulse responses using a partial Haar transform. *IEEE Transactions on Signal Processing* 53(3), 966–976 (2005)
6. Noskoski, O.A., Bermudez, J.C.M.: Wavelet-packet-based adaptive algorithm for sparse impulse response identification. In: Proc. of the ICASSP., vol. 3, pp. 1321–1324 (2007)
7. ITU-T Recommendation G. 168: Digital Network Echo Cancellers (2004)
8. Mallat, S.: A wavelet tour of signal processing. Academic Press, London (1998)
9. Sayed, A.H.: Fundamentals of Adaptive Filtering. Wiley, Chichester (2003)

Simulation Method of SAR Raw Echo for Urban Scene

Yingxi Zheng, Bo Zhou, Zhulin Zhong, and Ming Lv

Research Institute of Electronic Science and Technology of
University of Electronic Science and Technology of China

Chengdu, Sichuan Province, China
zhengyingxi@uestc.edu.cn

Abstract. In this paper, a SAR raw data simulation method for urban scene is presented based on the hypothesis that urban area is a set of vertical buildings placed over a random rough dielectric terrain. Facet model and Kirchhoff approach appropriately including multiple-scattering effects are adopted here to compute scattering coefficients in the scattering model, which operates in two-dimensional Fourier transformed domain. Methods of computing the scattering coefficients in different conditions are discussed in detail. Subsequently, the computational formulas and steps are also provided. It is known that the scattering model can effectively simulate the urban scene. The proposed simulation method of SAR raw echo turns out to be valid through simulation and analysis of the imaging to raw echo.

Keywords: SAR; facet; raw data simulation.

1 Introduction

SAR, one of the advanced techniques of radar imaging was developed in 1970s, which is applied widely in many areas such as military, ocean, agriculture and so on. Software simulation, which produces simulative echo and images, is a very important and economical method in research of SAR systems^[1].

Franceschetti^[2] shows a ground scene based digital elevation model (DEM) by establishing SARAS simulation system. A computing method of backscattered field of building scenes and some formulas are also provided in [3].

In this paper the methods of facet model and Kirchhoff approach are used to compute scattering coefficient based on the analysis to backscattering field and backscattering coefficient based upon [3].

2 Scattering Model of Building

Urban scene is supposed as a set of vertical buildings on a random rough dielectric terrain here. In case of just an isolated cube building for convenient analysis, whose geometrical figure is showed in fig1. Its vertical surface is parallel with axis Z and the bottom side is Z=0.

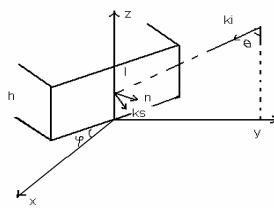


Fig. 1. Geometrical relation between building and ground

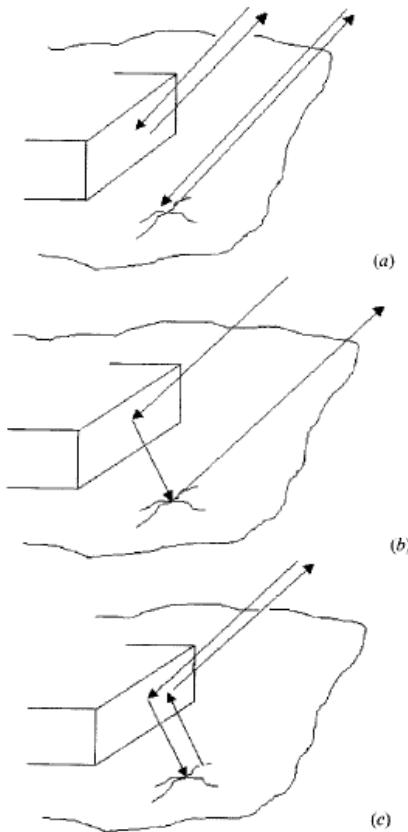


Fig. 2. (a) single, (b) double, and (c) triple scattering

In [4] the author analyzes the multiple scattering model and the conditions in point in detail, the relative conclusions of which are cited here directly. Superposition of first-, second-, and third-order contributions fully represent the scattered field; higher order mutual interactions do not give any contribution to the backscattered field to the radar antenna because the wall surface is supposed flat. In order to account for multiple scattering between buildings and terrain, we use Geometric optics (GO) to evaluate the field

reflected by the smooth wall toward the ground (first bounce) or the sensor (second or third bounce), and PO (Physical optics) or GO (depending on ground surface roughness) to evaluate the field scattered by the ground toward the wall (first or second bounce) or the sensor (second bounce).

In the prevenient literature there are many researches about the computing method of backscattering field which has only single scattering. The method in [2] is employed in the paper. The statistic characteristic of backscattering coefficient of single scattering and double scattering has been given in [3] (Table III, Table IV). Below the method of computing scattering coefficient in double and triple scattering based upon single facet will be deduced. The single-[fig.2 (a)], double-[fig.2 (b)], triple-[fig.2(c)] scattering are illuminated as follows.

3 Facet Model

Facet Model is physical model of simulating ground scattering characteristics. The natural ground scene can be described by amounts of small facets which are tangent to the surface of scenes. The electromagnetic field scattering characteristics of the ground scene result from coherent superposition of backscattering field of facet^[5]. The electromagnetic field scattering characteristics of facet is decided according to the roughness degree^[6] of the facet surface and medium material^[7]. The dimension of facet is bigger than the signal wavelength and less than resolution cell of radar so far^[3].

4 The Computation Method of Scattering Coefficient

4.1 Scattering Coefficient S_{wg} from Walls to the Ground

Because the relative is not backscattering between the direction of incident wave and scattering wave, namely $\bar{k}_s \neq -\bar{k}_i$, and the facet scattering coefficient computation formula can be not used directly here in literature [3] and need to be deduced again.

According to Kirchhoff approach theory scattering field physical optics of each facet can be deem approximately:

$$\bar{E}_s(\bar{r}) = \frac{jke^{jkr}}{4\pi r} E_0(I - \bar{k}_s \bar{k}_s) \cdot \bar{F}(\bar{k}_i, \bar{e}_i, \bar{n}) \cdot I_s \quad (1)$$

Wherein

$$\begin{aligned} \bar{F}(\bar{k}_i, \bar{e}_i, \bar{n}) = & -(\bar{e}_i \cdot \bar{q}_i)(\bar{n} \cdot \bar{k}_i)\bar{q}_i(1 - R_p) + (\bar{e}_i \cdot \bar{p}_i)(\bar{n} \times \bar{q}_i)(1 + R_q) \\ & + (\bar{e}_i \cdot \bar{q}_i)(\bar{k}_s \times (\bar{n} \times \bar{q}_i))(1 + R_p) + (\bar{e}_i \cdot \bar{p}_i)(\bar{n} \cdot \bar{k}_i)\bar{k}_s \times \bar{q}_i(1 - R_q) \end{aligned} \quad (2)$$

where \bar{r} is the vector from antenna to facet, k is wavenumber, I is unit matrix, $\bar{E}_s(\bar{r})$ is scattering field strength of facet along the direction of $\bar{k}_s, \bar{q}_i = \frac{\bar{k}_i \times \bar{n}}{|\bar{k}_i \times \bar{n}|}$,

$\bar{p}_i = \bar{q}_i \times \bar{k}_i$, R_p and R_q are respectively polarized Fresnel reflection coefficients along the direction of \bar{p}_i and \bar{q}_i . \bar{F} reflects the electromagnetism characteristics of polarization and facet surface and I_s reflects the shape of facet^[2].

Suppose horizontal polarization direction unit vectors of incident field and scattering field are $\bar{e}_{ih} = \frac{\bar{z} \times \bar{k}_i}{|\bar{z} \times \bar{k}_i|}$ and $\bar{e}_{sh} = \frac{\bar{z} \times \bar{k}_s}{|\bar{z} \times \bar{k}_s|}$ respectively, and the vertical polarization direction unit vectors are $\bar{e}_{iv} = \bar{e}_{ih} \times \bar{k}_i$ and $\bar{e}_{sv} = \bar{e}_{sh} \times \bar{k}_s$.

According to the geometrical relative in fig.1,

$$\bar{k}_i = -\sin \theta \bar{y} - \cos \theta \bar{z}, \bar{n} = \sin \varphi \bar{x} + \cos \varphi \bar{y} \quad (3)$$

Meantime, the scattering field strength along the lens reflection direction of incident wave is biggest because the surface of walls is flat. Here \bar{k}_s is reflection direction and

$$\bar{k}_s = \sin \theta \sin 2\varphi \bar{x} + \sin \theta \cos 2\varphi \bar{y} - \cos \theta \bar{z} \quad (4)$$

The elements of scattering matrix

$$S_{wg} = \begin{pmatrix} S_{hh} & S_{vh} \\ S_{hv} & S_{vv} \end{pmatrix} \quad (5)$$

can be given,

$$S_{hh} = \frac{2 \sin \theta \cos \varphi (\cos^2 \theta \cos^2 \varphi R_p - \sin^2 \varphi R_q)}{\cos^2 \theta \cos^2 \varphi + \sin^2 \varphi} \quad (6)$$

$$S_{hv} = \frac{2 \sin \theta \cos \theta \sin \varphi \cos^2 \varphi (R_p + R_q)}{\cos^2 \theta \cos^2 \varphi + \sin^2 \varphi} \quad (7)$$

$$S_{vh} = \frac{-2 \sin \theta \cos \theta \sin \varphi \cos^2 \varphi (R_p + R_q)}{\cos^2 \theta \cos^2 \varphi + \sin^2 \varphi} \quad (8)$$

$$S_{vv} = \frac{2 \sin \theta \cos \varphi (-\sin^2 \varphi R_p + \cos^2 \theta \cos^2 \varphi R_q)}{\cos^2 \theta \cos^2 \varphi + \sin^2 \varphi} \quad (9)$$

As $I_s = \int_s \exp[j(\bar{k}_i - \bar{k}_s) \cdot \bar{r}'] d\bar{r}'$ and $\bar{k}_i - \bar{k}_s$ are vertical to facet, $I_s = A$ and

A is the area of facet.

4.2 Scattering Coefficient in Other Conditions

4.2.1 Scattering Coefficient S_{wr} from Walls to Radar

It is similar to the condition from walls to the ground. Now let $-\bar{k}_i$ instead of \bar{k}_i , $-\bar{k}_i$ instead of \bar{k}_s , \bar{e}_{sh} instead of \bar{e}_{ih} , \bar{e}_{ih} instead of \bar{e}_{iv} and bring them into formula (2). The scattering matrix can be computed. Meantime, $I_s = A$ and A is the area of facet.

4.2.2 Scattering Coefficient S_{gw} from the Ground to Walls

There are two cases:

One: The scattering is backscattering when it is the second bound in triple scattering.

Two: When it is the first bound in double scattering, the normal vector of facet

4.2.3 Scattering Coefficient S_{gr} from the Ground to the Sensor

There are three cases:

One: When it is the second bound in double scattering, it can be computed by equation (2) directly.

Two: When it is the second bound in triple scattering, it belongs backscattering.

Three: When it is the first bound, it belongs backscattering.

4.3 Computation of Multiple-Scattering Coefficients

In conclusion from that, the computation of double scattering and triple scattering matrixes are as follows,

$$S = S_{gwr} + S_{wgr} = S_{wr} S_{gw} + S_{gr} S_{wg} \quad (10)$$

$$S = S_{wgwr} = S_{wr} S_{gw} S_{wg} \quad (11)$$

5 Echo Generation Model

We consider that echo is get by targets' backscattering coefficient through a system function $h(x, r)$, namely $s(x, r) = \sigma(x, r) \otimes h(x, r)$, and $h(x, r)$ is expressed^[8].

$$h(x, r) = w_a(x) w_r(r) \text{rect}\left\{\frac{2[r - R(x)]}{C}\right\} e^{j\frac{4\pi[r - R(x)]}{\lambda}} e^{j\frac{4\pi K[r - R(x)]^2}{C^2}}. \quad (12)$$

Where $\sigma(x, r)$ is backscattering coefficient, $w_a(\cdot)$ is azimuth antenna pattern, $w_r(\cdot)$ is distance antenna pattern, $\text{rect}(\cdot)$ is rectangular window function, $R(x)$ is the distance

between targets and radar, C is light velocity, λ is wavelength and K is FM-ratio of line frequency modulation signal (LFM).

The flow of echo simulation is as follows.

One: Generate building scene DEM data under background of Gauss random distribution.

Two: Compute the first scattering coefficient and the mean of scattering coefficient of the same distance cells to the ground facet.

Three: Compute the second scattering coefficient and the mean of scattering coefficient of the same distance cells to the ground facet and building facet.

Four: Compute the third scattering coefficient and the mean of scattering coefficient of the same distance cells to the ground facet and building facet.

Five: Perform multiplication between 2-D FFT of scattering coefficient $\sigma(x, r)$ and 2-D FFT of system function $h(x, r)$, which the result is $S(k, f)$.

Six: The raw echo $s(x, r)$ is provided by operating 2-D FFT to $S(k, f)$

6 Simulation Results

Table 1. The system parameters of airborne SAR simulation

Parameter	Value
Height	20000m
Wavelength	0.03m
Velocity of airplane	250m/s
Horizontal wave width	4.3°
Depression angle	25°
Vertical wave width	7.3°
Bandwidth	40MHz
Sampling rate	50MHz
Resolution	5m×5m

Table 2. The parameters of ground scene

Parameter	Value
Building height (h)	25m
Building width (b)	25m
Building length (l)	25m
Angle between wall and flight direction (φ)	0°
building and ground dielectric constant ϵ_r	4
ground height obeys gauss random distribution and coarseness degree: $k\sigma$	6.28
polarization mode	VV

The scattering figures are fig.3, fig.4, fig.5, fig.6 and image of raw signal is fig.7.

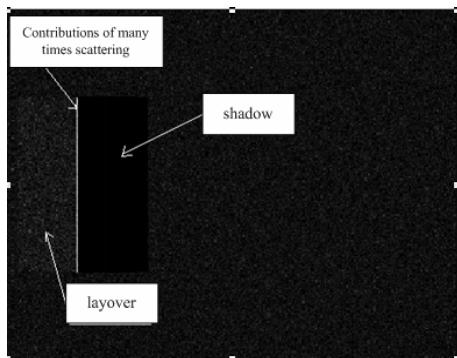


Fig. 3. Scattering figure when $h=40m$, $b=40m$, $l=25m$, $\varphi = 0^\circ$

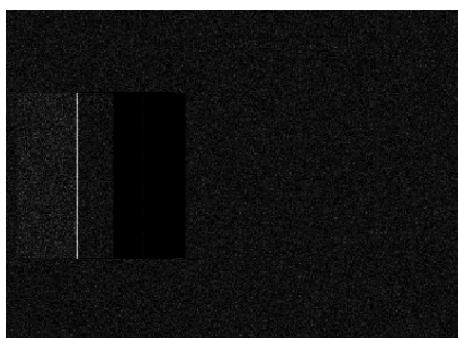


Fig. 4. Scattering figure when $h=40m$, $b=40m$, $l=40m$, $\varphi = 0^\circ$

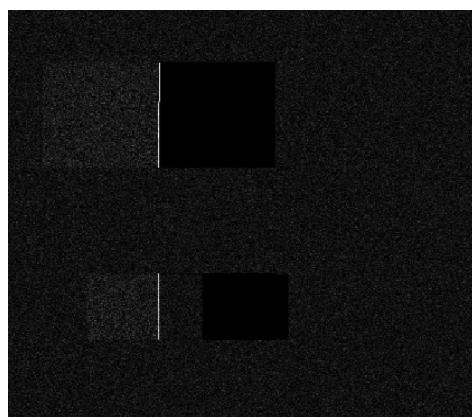


Fig. 5. Building scattering figure when $h=40m, b=40m, l=25m$, $\varphi = 0^\circ$ and $h=25m, b=25m, l=25m$, $\varphi = 0^\circ$

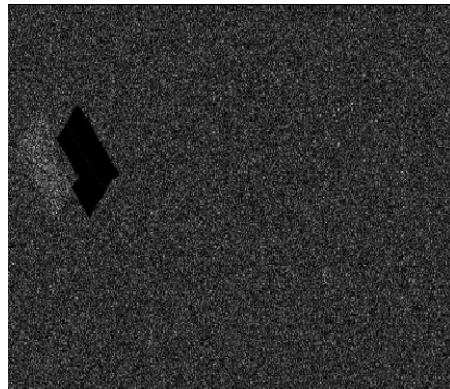


Fig. 6. Scattering figure when $h=40m, b=40m, l=40m, \varphi = 45^\circ$

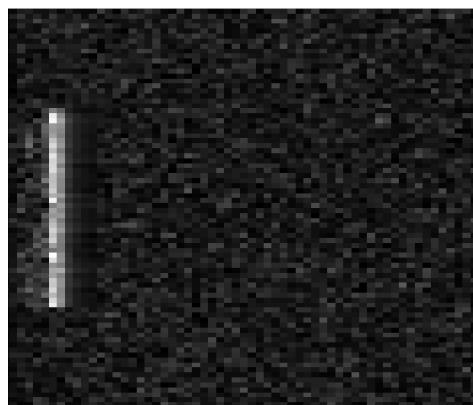


Fig. 7. Image when $h=40m, b=40m, l=25m, \varphi = 0^\circ$

From fig.2 and fig.3, we can obviously see layover from left to right, subsequently light scattering lines resulting from multiple scatterings of walls and shadow and special phenomenon of SAR image. Meanwhile, the locations of layover and shadow are relative to the length of the building. We can see from fig.7 that image is approach to corresponding scattering coefficient in fig. 3.

7 Conclusion

The paper gives the formulas of computing scattering coefficient from wall to ground in double and triple scattering based on facet building according to Kirchhoff approach and facet model. At the same time, the scattering coefficients and echoes of building scenes with different parameters are also simulated. The result of test shows that the simulative method can response the special characteristics of SAR image, such

as layover, shadow, azimuth base point and so on. Moreover, the method can simulate raw signals adapted various SARs to different roughness degrees, different dielectrics, different building shapes, different polarizations and so on.

Acknowledgments. The authors thank for Li Yangyuan' helpful advices.

References

1. Jinrao, S., Hong, S.: Simulation for natural scene SAR image. Rada science and technique 1(4), 200–214 (2003)
2. Franceschetti, G., Migliaccio, M., Riccio, D.: SARAS: a synthetic aperture radar (SAR) raw signal simulator. IEEE Trans, Geoscience and Remote Sensing 30(1), 110–123 (1992)
3. Franceschetti, G., Iodice, A., Riccio, D.: A canonical problem in electromagnetic backscattering from buildings. IEEE Trans, Geoscience and Remote Sensing 40(8), 1787–1801 (2002)
4. Franceschetti, G., Iodice, A., Riccio, D., et al.: SAR raw signal simulation for urban structures. IEEE Trans, Geoscience and Remote Sensing 41(9), 1986–1995 (1986)
5. Feng, W.: Raw Data Simulation for DistributionTargets of Space-borne SAR, mater degree paper (2003)
6. Jianchao, F., et al.: Microwave Remote Sensing-Radar Remote Sensing and Scattering and Radiation Theory of Face Targets. Science Press, BJ (1987)
7. Jie, C., Yingqing, Z., Chunsheng, L.: Research of simulation method for natural ground scene of space-borne SAR. Electronic transaction 29(9), 1202–1205 (2001)
8. Yongtan, L., et al.: Radar Imgng Technique, Ha Er Bin Industry Universtity Press, 10 (1990)

Underwater Passive Location Technology Using Dummy Time-Reversal Mirror

Xueli Sheng, Xizhong Bao, Junying Hui, and Guolong Liang

College of Underwater Acoustic Engineering, Harbin Engineering University,
HLJ, P.R. China
Shengxueli@yahoo.com.cn

Abstract. Time reversal (TR) refocuses the incident acoustic field back to the origin of a probe signal regardless of the complexity of the medium. However, the operation of a TR requires a cooperative probe source. This is impossible in passive work mode. In this paper, a dummy time-reversal mirror (DTRM) method is discussed to introduce the concept of time-reversal into underwater passive location. Several ocean acoustics experiments demonstrating the implementation of DTRM have been conducted. The proposed method is useful while at long source-receive ranges, or while object is moving.

Keywords: Dummy Time-reversal mirror, Passive location, adaptive refoule.

Introduction

Ideal channel can transform information invariably but underwater acoustic channel cannot because it's multipath character. The focusing of acoustic waves in inhomogeneous media is a common and difficult problem.

The concept of an acoustic time reversal mirror was introduced by Fink in 1989 as an extension of the concept of optical phase-conjugate mirror [1]. In 1991, Jackson and Dowling presented the idea of adapting the techniques of optical phase-conjugate mirrors to problems in underwater acoustics [2]. An experiment conducted by Kuperman in the Mediterranean Sea in April 1996 demonstrated that a time-reversal mirror can be implemented to spatially and temporally refocus an incident acoustic field back to its origin [3]. Over the last twenty years, acoustic time-reversal mirror have been demonstrated in the ocean and now the important application of time-reversal is in underwater communications, water tunnel acoustic measurements, reverberation nulling, etc[4],[5].

The use of time-reversal in the field of underwater passive location is a relatively new field. Generally, the typical time-reversal experiment needs a source to emit signal and a source array to re-emit time reversed signal. It is hard to be satisfied while locating a passive source. In this study, passive location technology based on an acoustic dummy time-reversal mirror (DTRM) is investigated. Instead of using a source array to re-emit reversed signal, model of Coherent Multipath Channels is established and the method makes use of established channels instead of actual ocean medium. Time reversed signal is convoluted with established channels by computer,

but not retransmitted to ocean. In the next section, TRM theory is reviewed. In Sec. 2, the proposed DTRM method is described and simulative results are given. In Sec. 3, experimental geometry and experiment results are described. Conclusions are drawn in Sec. 4.

1 Background Theory

The theory of time-reversal has already been presented [2],[6],[7]. Here we briefly review salient issues using the basic geometry of Kuperman's TRM experiment (shown schematically in Fig.1).

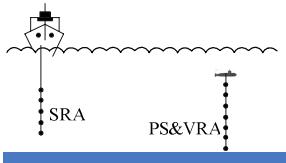


Fig. 1. Kuperman's experimental setup of TRM

A probe source (PS) indicated by one of the rectangles on the vertical receive array (VRA), sends out a pulse that is received at the source-receive array (SRA). The dispersed signal with all its multipath structure is time reversed and retransmitted by the SRA. The resulting signal multi-path structure collapses to a spatial and temporal focus at the original PS position that is co-located in range with the VRA.

Underwater acoustic channel can be regarded as a coherence and multipath channel while its spatial and temporal randomization is neglected [8]. We demonstrate the temporal and spatial adaptive focusing character of TRM from the point of view of ray theory.

Define $R_i(\omega)$ as the Fourier transform (FT) of $r_i(t)$, the signal received by the i th SRA receiver element. $R_i(\omega)$ can be described as the product of $S(\omega)$, the FT of the signal emitted by PS, and $H_i(\omega)$, channel's transform function from PS to the i th receiver element of the SRA

$$R_i(\omega) = H_i(\omega) \cdot S(\omega) . \quad (1)$$

Define $X_i(\omega)$ as the complex conjugate of $R_i(\omega)$, which is equal to time reverses $r_i(t)$ at time-domain.

$$X_i(\omega) = R_i^*(\omega) = H_i^*(\omega) \cdot S^*(\omega) . \quad (2)$$

Both side of Eq. (2) multiplied by $H(\omega)$, channel's transform function from the i th SRA element to somewhere in the ocean. The solution is given by

$$Z(\omega) = H(\omega) \cdot X_i(\omega) = H(\omega) \cdot H_i^*(\omega) \cdot S^*(\omega) . \quad (3)$$

$Z(\omega)$ can be regarded as the output of the TRM. Eq. (3) can be used to show that the TRM produces focusing in time as well as in space. Examine the TRM output at the focus point, that is, takes $H(\omega) = H_i(\omega)$. “ $H_i(\omega) \cdot H_i^*(\omega)$ ” in Eq. (3) is a real, even and positive function where the maximum of correlation can be got when IFT executed at $t=0$. This operation is matched filtering. The sum over SRA elements is a form of spatial matched-field processor.

The principle demonstrated upwards is the basic of acoustic time-reversal mirror in a range-independent waveguide.

2 Dummy Time-Reversal Mirror (DTRM)

2.1 Theory on Passive Location Using a Dummy Time-Reversal Mirror

As described above, TRM works as a form of matched filtering.

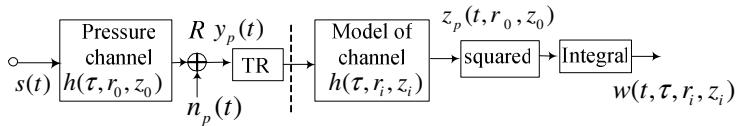


Fig. 2. Principle block diagram of DTRM

Principle block diagram of DTRM is shown in Fig. 2. Here, $s(t)$ is the noise radiated by passive point source located at $a_0(r_0, z_0)$. $h(\tau, r_0, z_0)$ is channel's transform function from passive source to receiver R and $y_p(t)$ is receiver's output that ambient noise be taken into account. A mathematical model of coherent multipath channel from receiver to every interested point $a_i(r_i, z_i)$ (include a_0) in the ocean, which marked as $h(\tau, r_i, z_i)$, can be established by using sound speed, depth of sea area and sea bottom character measured practically. The received signal $y_p(t)$ is time reversed and ‘retransmitted’ by computer and the resulting signal $z_p(t, r_i, z_i)$ is an equivalence of the TRM reslut at the position $a_i(r_i, z_i)$. In this course, $y_p(t)$ dose not retransmitted to ocean factually but convoluted with channel model $h(\tau, r_i, z_i)$ by computer. $z_p(t, r_i, z_i)$ can be regarded as IFT of $Z(\omega)$ shown in Eq. 3, which multipath structure collapses at the original passive point source location $a_0(r_0, z_0)$, that is mean the energy $w(t, \tau, r_0, z_0)$ of $z_p(t, r_0, z_0)$ is maximum of all $w(t, \tau, r_i, z_i)$. This is the concept of passive location by DTRM.

2.2 Model of Coherent Multipath Channel

In this paper, a mathematical model of coherent multipath channel is established by using the theory of ray acoustic and the method of eigenrays to simplify the physical

of underwater sound channel. Multipath channel's system function is determined by eigenrays which are sound rays reaching the receive point. The sea surface and sea bottom has a profound effect on propagation in most applications of underwater sound where source or receiver lie at shallow depth. Above all, establishing coherent multipath channel model consists of two cruces. First, search all the important eigenrays quickly and correctly. Secondly, estimate acoustic parameters of sea surface and sea bottom precision accurately [9].

The method of establishing a model of coherent multipath channel based on the theory of ray acoustic and eigenrays is described in [9].

A criterion for the roughness or smoothness of a surface is given by the *Rayleigh parameter* [10] and the average coefficient of sea surface is related to frequency by

$$V_s = 1 - 0.45 \cdot (f \cdot H_{\text{avg}})^{3/2} \cdot \cos \theta_0 . \quad (4)$$

where f is the frequency in kilohertz, and H_{avg} is wave's average height in meter.

Acoustic parameters of sea bottom is estimated by using the three-parameters model [11]

$$-\ln|V_b(\theta)| = \begin{cases} Q \cdot \theta & 0 < \theta < \theta^* \\ -\ln|V_{b0}| = \text{const} & \theta^* < \theta < \frac{\pi}{2} \end{cases} . \quad (5)$$

A critical angle θ^* exists such that complete reflection occurs at grazing angles θ less than critical. Q is the slope of sea bottom reflection at grazing angles less than θ^* and V_{b0} is sea bottom reflection at grazing angles more than θ^* . Description of the three-parameters model is shown in Fig. 3.

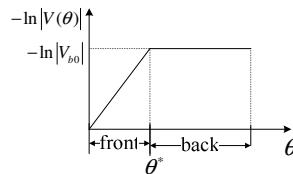


Fig. 3. Three-parameters model

The amplitude A , time delay τ , and grazing angle θ of eigenrays reaching a given point in the sea can be got from model of coherent multipath channel and the pressure channel from this point to source can be described as

$$h_p(x, y) = \sum_{i=1}^N A_i \delta(t - \tau_i) . \quad (6)$$

where N is the number of eigenrays.

2.3 Simulation Result

For the purpose of simulation, the source position $(r_s, z_s) = (1000m, 20m)$ and the receiver position $(r_0, z_0) = (0m, 0m)$ is used. Sound speed profile used to model the channel is shown in Fig. 4(a). Hypothetically, channel modeled from passive source to receiver is equal to actual channel. SNR is defined at output of filter. Location result of one target by DTRM is shown in Fig. 4(b), while $SNR = -4dB$ and bandwidth from 2.5kHz to 7.5kHz over a 0.5-s time window.

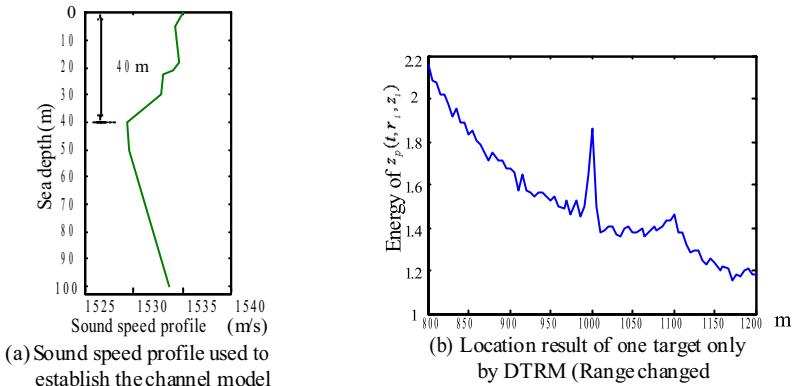


Fig. 4. Simulation result of DTRM

In Fig. 4(b), x-axis denotes range r_i of observer and y-axis denotes DTRM's energy output at observer. Simulation result proves the passive location ability of TRM by one sensor.

Three-dimensional location result by DTRM is shown in Fig. 5, in which the source position $(r_s, z_s) = (1000m, 40m)$ relative to the receiver. Color in the figure shows the energy output of DTRM.

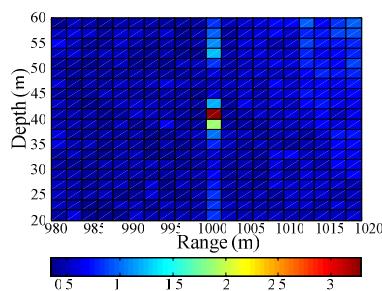


Fig. 5. The DTRM refocused energy distribution in shallow sea

The peak value of energy output is present at range 1000m to 1020m, depth 40m to 42m, which mean that the source can be located passively by the DTRM refocused energy distribution.

3 DTRM Experiment

An experiment conducted in the ocean in 2006 demonstrated that a DTRM can be implemented to locate a passive source. Fig. 6(a). is a schematic of the experiment. The DTRM was implemented by one receive sensor in 160-m-deep water. An underwater explosive sound source was used as a passive source which distance from receive sensor was 9.4 km measured by GPS.

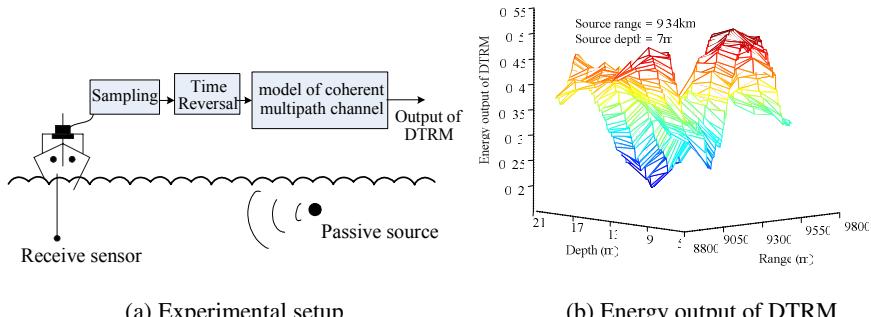


Fig. 6. DTRM experiment in the ocean

Energy output of DTRM is shown in Fig. 6(b). The waterfall plot shows the energy output of DTRM as a function of depth and range. Maximum energy strength of dummy propagation using data got by receive sensor indicates the source range at 9.3 km and depth at 7m, which accord with the result got by GPS. This experiment demonstrated that a DTRM can be implemented to refocus a passive source back to its origin in the ocean.

The second experiment was performed off the Songhua lake of China in September 2007 as indicated in Fig. 7. A three-vector-buoy acoustic positioning array was used

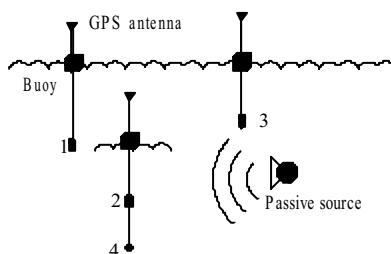


Fig. 7. Experimental setup of DTRM experiment in the lake

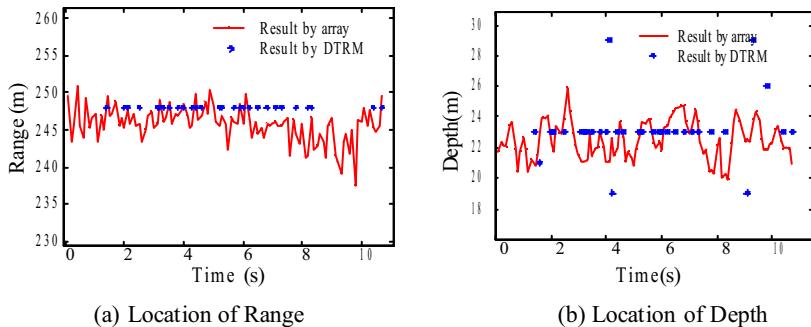


Fig. 8. Location results of experiment in the lake

to operate passive source's three-dimension location. 1, 2 and 3 indicated in Fig. 7 were vector sensors in 16-m, 8-m and 14-m deep, and 4 was press sensor in 22-m deep. The DTRM was implemented by sensor 2. The bandwidth of passive source is from 800Hz to 2000Hz.

Locating result obtained by buoy array is shown in Fig. 8 by line and DTRM by dot.

Compared with classically passive location method using buoy array, DTRM also can locate passive source accurately merely several inefficacy refocus appears during the process.

The third experiment in the ocean in April 2006 demonstrated that DTRM with one receiver can be implemented to locate a passive moving source as well.

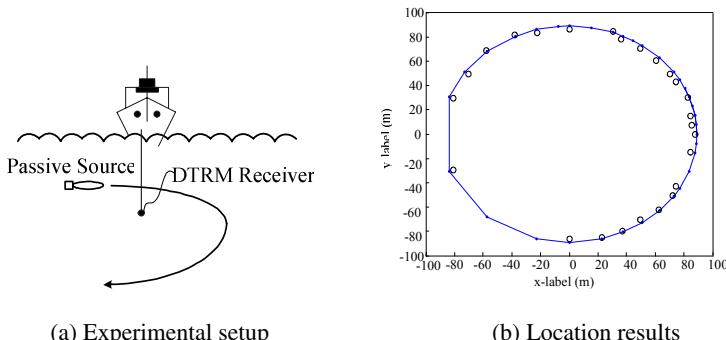


Fig. 9. Experiment of DTRM to locate a passive moving source

Fig. 9 (a) is a schematic of the experiment. Receiver is 5-elements vector array and DTRM receiver indicated by one of them. Passive source move around the array and two- dimension's location result obtained by using array is shown in Fig. 9 (b) as the lines and DTRM' result as dot. This may be a result that DTRM can be implemented to locate a passive moving source.

4 Conclusion

The concept of the Dummy Time Reversal Mirror is introduced. This is used to locate passive point source in the ocean. We have implemented acoustic DTRM in the ocean and hence demonstrated that DTRM is realizable in ocean using a receive sensor and rather simple signal processing. We also have investigated that DTRM can be implemented to locate a passive moving source as well as it would to motionless source.

References

1. Fink, M., Prada, C., Wu, F., Cassereau, D.: Self focusing in homogeneous media with ‘time reversal’ acoustic mirror. In: Proceedings of the IEEE Ultrasonic Symposium 1989, Montreal, PQ, Canada, pp. 681–686 (1989)
2. Jackson, D.R., Dowling, D.R.: Phase Conjugation in underwater acoustic. *J. Acoust. Soc. Am.* 89, 171–181 (1991)
3. Kuperman, W.A., Hodkiss, W.S., Song, H.C.: Phase conjugation in the ocean: Experimental demonstration of an acoustic time-reversal mirror. *J. Acoust. Soc. Am.* 103(1), 25–40 (1998)
4. Edelmann, G.F., Akal, T., Kuperman, W.A.: An initial demonstration of underwater acoustic communications using time reversal. *IEEE J. Ocean. Eng.* 27, 602–609 (2002)
5. Song, H.C., Hodkiss, W.S., Kuperman, W.A., Sabra, K.G., Akal, T.: Passive reverberation nulling for target enhancement. *J. Acoust. Soc. Am.* 122(6), 3296–3303 (2007)
6. Jackson, D.R., Dowling, D.R.: Narrow-band performance of phase-conjugate arrays in dynamic random media. *J. Acoust. Soc. Am.* 91, 3257–3277 (1992)
7. Dowling, D.R.: Phase-conjugate array focusing in a moving medium. *J. Acoust. Soc. Am.* 94, 1716–1718 (1993)
8. Hui, J., Sheng, X.: Underwater Acoustic Channel, 2nd edn. Defense Industry Press, Beijing (2007)
9. Fan, M.: A Study on Simulation & Application of Underwater Sound Channel. A Dissertation for the Degree of D.Eng, Harbin Engineering University (2000)
10. Brekhovskikh, L.M.: Fundamentals of Ocean Acoustics. Science Press, Beijing (1983)
11. Wang, D., Shang, E.: Underwater Acoustics. Science Press, Beijing (1981)

Multifractal Analysis on the Sphere

Emilie Koenig and Pierre Chainais

LIMOS UMR 6158, University of Clermont-Ferrand II, 63170 Aubière Cedex, France
`emilie.koenig@isima.fr, pchainai@isima.fr`

Abstract. A new generation of instruments in astrophysics or vision now provide spherical data. These spherical data may present a self-similarity property while no spherical analysis tool is yet available to characterize this property. In this paper we present a first numerical study of the extension of multifractal analysis onto the sphere using spherical wavelet transforms. We use a model of multifractal spherical textures as a reference to test this approach. The results of the spherical analysis appear qualitatively satisfactory but not as accurate as those of the usual 2D multifractal analysis.

1 Introduction

Multifractal analysis is used to characterize the self-similarity property of objects. This analysis method has been used in various domains, each presenting self-similar data, such as turbulence in physics [1], network traffic [2], DNA series in biology [3] or in the study of natural images [4]. The development of new instruments in several domains, as astrophysics [5] or vision [6], leads to a new generation of data: spherical data. These new elements may present a self-similarity property but no multifractal analysis has yet been proposed to characterize it. In this paper, we introduce a spherical multifractal analysis based on spherical wavelets as a natural extension of the usual 2D analysis. A numerical study is carried out with help of synthetic spherical textures generated by an extension of the Compound Poisson Cascade (CPC) model [7] on the sphere. These numerical experiments show that the spherical analysis appear qualitatively satisfactory but not as accurate as those of the usual 2D multifractal analysis.

The paper is organized as follows. First, we recall on the 2D multifractal analysis theory and introduce its extension onto the sphere. Next, we present a family of multifractal processes on the sphere, namely the Compound Poisson Cascades, to be used as a reference to test the method. Then, we perform numerical experiments to test the spherical multifractal analysis. Finally, we comment on our results and discuss about possible forthcoming improvements.

2 Multifractal Analysis on the Sphere

2.1 2D Multifractal Analysis

The self similarity property [8] refers to the fact that a part of an object resemble the whole object itself. This property can be associated to some deterministic

geometrical feature: a dilated version of some part of a snowflake is similar to the entire snowflake. An object can also be self similar in a statistical sense. For example, a dilated segment of an Internet traffic signal is statistically similar to the whole signal. At least in some range no scale plays a specific role. Deterministic and stochastic fractals are examples of such objects. In the following, we will refer to an *object* for either a function or a realization of a stochastic process. Formally, the self similarity property is usually characterized by power law scalings, and more precisely by the evolution of partition functions $S_q(a)$ with scale a . For positive functions f , partition functions can be defined as estimates of the q th-order moments of the localized box averages for a covering set of N_a positions indexed by l :

$$S_q^{box}(a) = \frac{1}{N_a} \sum_{l=1}^{N_a} \varepsilon_a^q(l) \propto a^{\tau(q)} \text{ where } \varepsilon_a(l) = \frac{1}{a} \int_{l-a/2}^{l+a/2} \beta_0\left(\frac{x-l}{a}\right) f(x) dx , \quad (1)$$

where β_0 is a positive weighting function. In the simplest case, $\beta_0 = \mathbb{1}_{x \in [-1/2, 1/2]}$. The coefficients $\varepsilon_a(l)$ are called *aggregation coefficients* [9]. For more general functions, the partition functions are often defined as wavelet-based partition functions as given by [2]:

$$S_q^{swt}(a = 2^j) = \frac{1}{N_j} \sum_{l=1}^{N_j} |d_f(j, l)|^q \propto 2^{j\zeta(q)} , \quad (2)$$

where $d_f(j, l)$ are the L^1 normalized discrete wavelet coefficients of the object f under study and N_j is the number of wavelet coefficients at octave $j = \log_2(a)$. The function $\zeta(q)$ defines a set of so-called *multipactal exponents*. In the simplest case, $\zeta(q) = qH$ so that all the exponents are described by a single parameter H . The object is then called *monofractal*. A monofractal object is characterized by a unique fractal dimension. This is for instance the case of the fractional Brownian motion. In general, the function $\zeta(q)$ takes the form $\zeta(q) = qH + \tau(q)$ where the function $\tau(q)$ is a non-linear function such that $\tau(0) = \tau(1) = 0$ and describes the deviation of $\zeta(q)$ from a linear function. The object is then called *multipactal*. The function $\zeta(q)$ is then characterized both by $H = \zeta(1)$ and $\tau(q) = \zeta(q) - qH$. The function $\tau(q)$ features the multipactality of the object. The estimated function $\zeta(q)$ may also be used as an input set of parameters for a model such as the Compound Poisson Cascades described in Sect. 3.1.

2.2 Extension of the Multifractal Analysis to the Sphere

We propose to extend the multifractal analysis described above to the sphere by using the same formulas but with spherical wavelets. We use a continuous spherical wavelet transform. The multifractal analysis is then only valid for $q \geq 0$ otherwise it will be numerically unstable and will raise theoretical problems [10]. Different constructions of spherical wavelet transforms (SWT) have been proposed.

P. Schröder and W. Sweldens [11] wanted to extend the discrete wavelet transform to any manifolds and in particular to the sphere. They have used a subdivision grid on the sphere (see Fig. 1(a)). Such a grid is based on the iterative division of the faces of an icosahedron in 4 new faces of equal area. These authors have constructed spherical wavelet functions with help of the lifting scheme and used these functions for compression. Two softwares are based on this transform: SD (<http://www.multires.caltech.edu/software/sd/>) developed by P. Schröder in 1998 and itkSWaveletSource [12] developed by Y. Gao et al. in 2007.

The extension of the wavelet construction made by J.-L Starck et al. [13] was motivated by the study of the cosmological microwave background (CMB) in astrophysics. These data are therefore mapped on the HEALPix (Hierarchical Equal Area isoLatitude Pixelization, <http://healpix.jpl.nasa.gov>) grid [14] (see Fig. 1(b)). The meshes of this grid are distributed on isolatitudes and have the same area so that every mesh has the same weight. The spherical wavelet construction of J.-L Starck et al. based on the spherical harmonic transform is implemented in a software available from the MRS Home Page (<http://jstarck.free.fr/mrs.html>).

J.-P. Antoine et al. [15] have introduced a spherical wavelet transform based on a correspondence between the plane and the sphere. The translation and dilation used in the definition of the 2D wavelet transform are replaced respectively by the rotation and the stereographic dilation on the sphere. The rotation $R(\rho)$, where ρ is a set of Euler angles, and dilation D_a operators are then defined by

$$(R(\rho)f)(\omega) = f(R_\rho^{-1}\omega) . \quad (3)$$

$$(D_a f)(\omega) = \lambda(a, \theta)^{1/2} f(D_a^{-1}\omega) . \quad (4)$$

where $f(\omega)$ is a function of the spherical coordinates $\omega = (\theta, \phi)$. The stereographic dilation maps the sphere without its South Pole on the plane. It imposes the use of an equiangular grid (see Fig. 1(c)). The two major defaults of this grid are the oversampling of the poles and the area variation of meshes. However, we chose this tool because it is efficient and easy to use in practice. The spherical continuous wavelet transform of $f(\omega)$ is the convolution between the function and the rotated and dilated versions of the mother wavelet Ψ

$$W_\Psi^f(\rho, a) = \langle \Psi_{a, \rho} | f \rangle = \int_{S^2} d\mu(\omega) \overline{[R_\rho D_a \Psi](\omega)} f(\omega) . \quad (5)$$

Then, the spherical multifractal analysis will be performed by replacing $d_f(j, l)$ by $W_\Psi^f(\rho, a)$ in (2) for a discrete sequence of tangential scales $a_j = 2 \tan((\pi/4)2^{-j})$, $1 \leq j \leq J$. This wavelet transform is implemented in the Matlab toolbox YAWTb (Yet Another Wavelet Toolbox, developed by L. Jacques et al. in 2002, <http://rhea.tele.ucl.ac.be/yawtb/>) [16].

The multifractal coefficients $\zeta(q)$ can only be calculated using a L^1 normalized wavelet. In practice, if $\Psi^{(2)}$ is the L^2 normalized spherical wavelet function, the

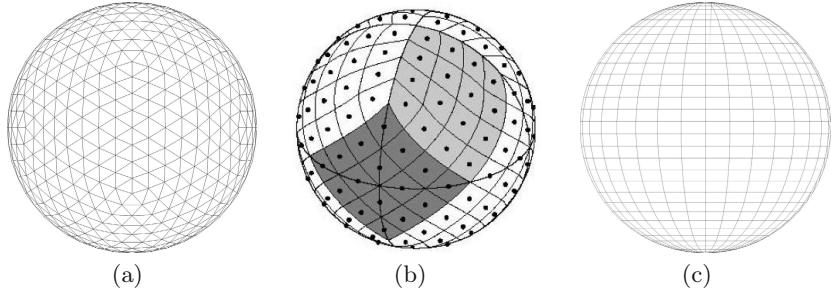


Fig. 1. (a) Subdivision grid [11]; (b) HEALPix grid [14]; (c) Equiangular grid [15]

L^1 normalized spherical wavelet function $\Psi^{(1)}$ is such as (see proposition 2.1 in [15])¹

$$\Psi(\theta, \phi) = \Psi^{(1)}(\theta, \phi) = \frac{\Psi^{(2)}(\theta, \phi)}{1 + \cos \theta} . \quad (6)$$

3 Compound Poisson Cascades (CPC) on the Sphere

A tunable self similar process generated on the sphere is used to validate the multifractal analysis described above. We have chosen the CPC model, a family of multifractal processes that can be defined in N dimensions [7] and in particular directly on the sphere (without any texture mapping), as explained below.

3.1 Compound Poisson Cascades in the Plane

A 2D CPC is a stochastic process used to synthesize gray level images. The pixel located at (x, y) takes the value given by

$$Q_\ell(x, y) = \frac{\prod_{(x_i, y_i, r_i) \in C_\ell(x, y)} W_i}{\mathbb{E} \left[\prod_{(x_i, y_i, r_i) \in C_\ell(x, y)} W_i \right]} . \quad (7)$$

where (x_i, y_i, r_i) result from a Poisson point process in the half space above the image plane. The coordinates x_i and y_i are uniformly distributed on the plane and the r_i ($\ell \leq r_i \leq 1$) are distributed with density $\propto 1/r_i^3$. The 3D points are weighted by i.i.d. random multipliers $W_i > 0$. The pixel value is computed as the product of the multipliers belonging to a cone pointing to this pixel (see Fig. 2(a)). In another interpretation of (7), the multiplier W_i can be seen as a light source influencing a part of the image. The size of its base is dilated depending on the height r_i of the multiplier (see Fig. 2(b)). Such models generate purely multifractal textures characterized by the non linear function $\tau(q)$. For instance, $S_q^{box}(a) \propto a^{\tau(q)}$. The distribution of the multipliers W_i prescribes $\tau(q)$ since

¹ We have implemented the L^1 spherical Mexican Hat wavelet in our YAWTb version.

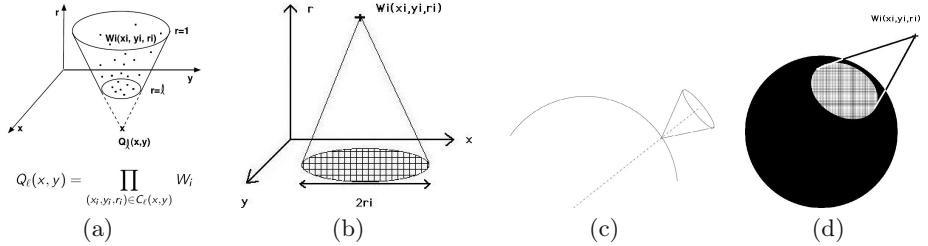


Fig. 2. IDC construction (a) and (b) in 2D; (c) and (d) on the sphere

$\tau(q) = q(\mathbb{E}W_i - 1) + 1 - \mathbb{E}W^q$. These very singular textures are often called *bare* textures ($H = 0$). More regular and smoother multifractal textures are obtained using a self-similar $1/\|k\|^H$ ($H > 0$) low-pass filter in the Fourier domain. The resulting texture is then characterized by the function $\zeta(q) = qH + \tau(q)$. Such filtered CPC are called *dressed* textures.

3.2 Extension of the Compound Poisson Cascades to the Sphere

The extension of CPC on the sphere uses the same principle as the 2D construction (see Fig. 2(c) and 2(d)). However, several choices are possible to define a dilation on the sphere, e.g. the stereographic dilation or the dilation of a solid angle. We have chosen this last dilation because it is the most direct equivalent of the dilation in the plane and it receives intuitive physical interpretation. The spherical texture (see Fig. 3(a)) is then generated by using (7), replacing the 2D coordinates by spherical ones (θ, ϕ) and adapting the distribution of scales r_i (not detailed here for brevity). As explained in Sect. 3.1, the resulting bare spherical texture, characterized by the function $\tau(q)$, has to be filtered to get a dressed texture characterized by $\zeta(q) = qH + \tau(q)$. This filtering is carried out in the (discrete) spherical harmonic domain [17] which is equivalent to the (discrete) Fourier domain. In brief, a function on the sphere $f(\theta, \phi)$ can be expanded on the set of spherical harmonics $Y_l^m(\theta, \phi)$ of degree l and order m

$$f(\theta, \phi) = \sum_{l \geq 0} \sum_{|m| \leq l} \hat{f}(l, m) Y_l^m(\theta, \phi) . \quad (8)$$

where the $\hat{f}(l, m)$ are the spherical harmonic coefficients of f . A convolution theorem exists on the sphere to describe azimuthally symmetric filters $\hat{h}(l, 0)$:

$$\widehat{(f * h)}(l, m) = 2\pi \sqrt{\frac{4\pi}{2l+1}} \hat{f}(l, m) \hat{h}(l, 0) . \quad (9)$$

Using the rough correspondence between the spherical indices (l, m) and the Fourier vector $\sqrt{l^2 + m^2} \sim k$ [18], we have chosen

$$\hat{h}(l, m) = \frac{1}{2\pi} \sqrt{\frac{2l+1}{4\pi}} \frac{1}{\sqrt{l^2 + m^2}^H} . \quad (10)$$

as the spherical equivalence of the $1/\|k\|^H$ filter. We know that this filtering is still approximative due to aliasing problems but it appears to be sufficiently precise since for this specific choice, aliasing effects are small. The resulting dressed texture can be characterized by a set of multifractal exponents $\zeta(q) = \tau(q) + qH$ (see Fig. 3(b)).

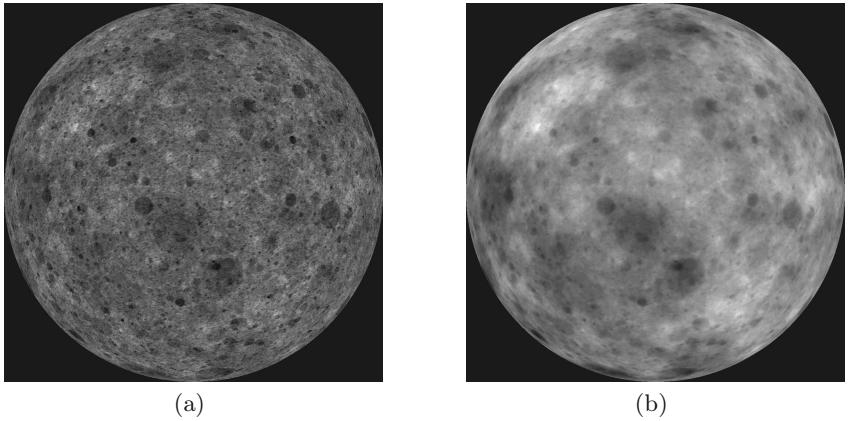


Fig. 3. CPC textures on the sphere: (a) bare texture characterized by $\tau(q)$ and (b) dressed texture characterized by $\zeta(q) = \tau(q) + qH$

4 Numerical Study

The multifractal analysis introduced in Sect. 2 is applied on a set of spherical textures generated with the model presented in Sect. 3. The spherical multifractal analysis will give an estimate $\hat{\tau}(q)$ of the exponents $\tau(q)$ prescribed in the synthesis of the CPC processes (see [7] for details). To evaluate the quality of the estimated $\hat{\tau}(q)$, we compare our results to the theoretical $\tau(q)$, to the $\hat{\tau}_{box}(q)$ using [11] for bare textures only and to the well-known results of the same experiments performed in the plane. The filtering using spherical harmonics is performed with the S2Kit (<http://www.cs.dartmouth.edu/~geelong/sphere/>) package [19] and the spherical wavelet transform is done with YAWTb. The wavelet used is the spherical Mexican Hat Wavelet, often picked up in astronomical applications of spherical wavelets [20].

Two sets of 30 spherical and 2D textures corresponding to the so-called log-exponential CPC with theoretical $\tau(q) = 1 - (1 + T)^q / (1 + qT)$ have been computed: bare textures with parameter $T = 0.7$ and the corresponding dressed textures with filter parameter $H = 0.55$. The choice $T = 0.7$ corresponds to very multifractal textures, that is very non linear $\tau(q)$.

As a first consistency check, we have estimated the multifractal coefficients using the aggregation coefficients method. This analysis can be performed in the direct space. This is quite difficult to implement because of the need to locate the grid vertices and the huge number of iterations. This implementation is also very expensive in computation time and is dependent on the spherical grid. Another implementation trades on the process used in YAWTb to compute wavelet coefficients. This method uses the spherical harmonic transform. It permits an efficient use of the aggregation coefficients mentioned in [1]. The results of the aggregation coefficients analysis are presented in Fig. 4. The exponents $\hat{\tau}_{box}(q)$ accurately estimate the theoretical $\tau(q)$ which validates the spherical CPC synthesis procedure.

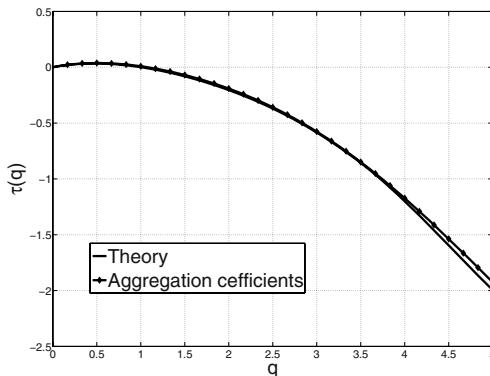


Fig. 4. Estimates based on aggregation coefficients $\hat{\tau}_{box}(q)$ for bare textures

Scaling exponents estimates are obtained from linear regressions in $\log_2 S_q(j)$ vs. $j = \log_2(a)$ diagrams over 5 octaves. The values of q are chosen in the range $0 \leq q \leq 5$. This range is restricted to positive values because the estimation for negative values is numerically unstable. Furthermore, estimates for $q \geq 5$ are expected to be either statistically inaccurate or uninformative.

Figure 5(a) presents the wavelet-based estimates $\hat{\tau}_{swt}(q)$ for the bare spherical textures compared to the theoretical values $\tau(q)$ and the results of the same experiments with flat 2D images. The corresponding spherical S_q^{swt} functions present a quite nice power law behavior onto a large range of scales and the behavior (trend and curvature) of $\hat{\tau}_{swt}(q)$ is quite similar to the theoretical $\tau(q)$. However, SWT-based estimates suffer from systematic bias in contrast with estimates based on box averages $\hat{\tau}_{box}(q)$ that were close to perfect. A noticeable defect of $\hat{\tau}_{swt}(q)$ is that it is not consistent with the fundamental property of bare CPCs $\hat{\tau}_{swt}(1) = -0.05 \neq 0$. The 2D equivalent wavelet analysis performed with planar 2D textures based on the same CPC model would give more consistent results.

The behavior (trend and curvature) of $\hat{\zeta}_{swt}(q)$ is quite similar to the theoretical $\zeta(q)$, see Fig. 5(b). However, we observe again that the $\hat{\zeta}_{swt}(q)$ systematically underestimate the expected $\zeta(q)$. As a consequence, the parameter H is badly

estimated since $\hat{H} = \hat{\zeta}_{swt}(1) = 0.46 < H = 0.55$. This bias is consistent with the bias previously observed on the $\hat{\tau}_{swt}(q)$ for bare textures. This indicates that the spherical $1/\|k\|^H$ filtering operation seems to work satisfactorily.

For some given model and using equivalent statistics, the variance of spherical estimates is of the same order of magnitude as the variance of 2D estimates (of the order of 5%, not represented here for sake of clarity).

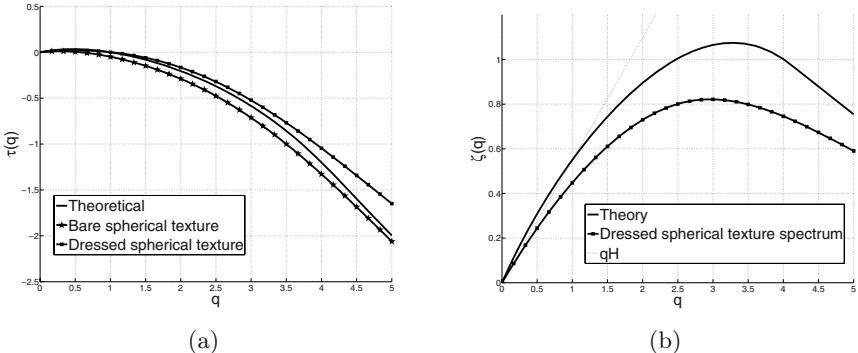


Fig. 5. (a) SWT $\hat{\tau}_{swt}(q)$; (b) Multifractal analysis of dressed textures, $\hat{\zeta}(q)$ of the form $qH + \tau(q)$

5 Conclusions and Perspectives

The purpose of this work was to study the relevance of a natural generalization of the wavelet based multifractal analysis from the usual 2D cartesian space to the sphere. Such a tool would be useful to characterize the scale invariance of spherical data. We have presented a numerical study of this approach based on the use of the CPC stochastic processes with prescribed multiscaling properties. These processes can be synthesized directly on the sphere without any mapping artifact. For better efficiency, we have implemented the multifractal analysis based on aggregation coefficients using the YAWTb wavelet transform process. This analysis confirms that the synthetic processes have the prescribed properties indeed. Despite their ability to capture the main behavior of the scaling exponents, the wavelet based estimates $\hat{\tau}_{swt}(q)$ and $\hat{\zeta}_{swt}(q)$ are not as accurate as their usual 2D equivalent. Note that many intricate problems arise from the spherical geometry. Indeed, we can not exclude that the chosen mesh grid (equiangular, subdivision, HEALPix,...) have some influence. In particular, the equiangular grid is not a multiresolution grid. This problem is combined to the use of continuous spherical wavelets which doesn't form the desired multiresolution basis commonly used in 2D. This preliminary work shows that the extension of multifractal analysis on the sphere using spherical wavelets may work but is not so immediate. Forthcoming work is needed to get more accurate estimates.

References

1. Frisch, U.: *Turbulence, The legacy of A. Kolmogorov*. Cambridge University Press, Cambridge (1995)
2. Abry, P., Flandrin, P., Taqqu, M., Veitch, D.: Wavelets for the analysis, estimation and synthesis of scaling data. In: *Self-Similar Network Traffic and Performance Evaluation*, Wiley-Interscience, Chichester (2000)
3. Arneodo, A., Bacry, E., Graves, P.V., Muzy, J.F.: Characterizing long-range correlations in dna sequences from wavelet analysis. *Physical Review Letters* 74(16), 3293–3296 (1995)
4. Turiel, A., Perez-Vincente, C.J., Grazzini, J.: Numerical methods for the estimation of multifractal singularity spectra on sampled data: a comparative study. *Journal of Computational Physics* 216(1), 362–390 (2006)
5. Marinucci, D., Pietrobon, D., Balbi, A., Baldi, P., Cabella, P., Kerkyacharian, G., Natoli, P., Picard, D., Vottorio, N.: Spherical needlets for cosmic microwave background data analysis. *Mon. Not. R. Astron. Soc.* 383(2), 539–545 (2007)
6. Tosic, I., Bogdanova, I., Frossard, P., Vanderghenst, P.: Multiresolution motion estimation for omnidirectional images. In: *EUSIPCO* (2005)
7. Chainais, P.: Infinitely divisible cascades to model the statistics of natural images. *IEEE Trans. on Pattern Anal. Mach. Intell.* 29(1) (2007)
8. Mandelbrot, B.: *The fractal geometry of Nature*. W.H. Freeman and Co, New York (1982)
9. Lashermes, B., Abry, P., Chainais, P.: New insights on the estimation of scaling exponents. *Int. J. of Wavelets, Multiresolution and Information Processing* 2, 497–523 (2004)
10. Jaffard, S.: Multifractal formalism for functions. *S.I.A.M.* 28(4), 944–998 (1997)
11. Schröder, P., Sweldens, W.: Spherical wavelets: Efficiently representing functions on a sphere. In: *Computer Graphics Proceedings SIGGRAPH 1995*, vol. 29, pp. 161–172 (1995)
12. Gao, Y., Nain, D., LeFaucheur, X., Tannenbaum, A.: Spherical wavelet itk filter. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) *MICCAI 2007*, Part I. LNCS, vol. 4791, Springer, Heidelberg (2007)
13. Starck, J.L., Moudden, Y., Abrial, P., Nguyen, M.: Wavelets, ridgelets and curvelets on the sphere. *Astron. astrophys.* 446, 1191–1204 (2006)
14. Gorski, K., Hivon, E., Banday, A., Wandelt, B., Hansen, F., Reinecke, M., Bartelmann, M.: Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *Astrophys. J.* 622(2), 759–771 (2005)
15. Antoine, J.P., Demanet, D., Jacques, L., Vandergheynst, P.: Wavelets on the sphere: implementation and approximations. *Appl. Comp. Harmon. Anal.* 13, 177–200 (2002)
16. Jacques, L.: *Ondelettes, repères et couronne solaire*. PhD thesis, Université catholique de Louvain (2004)
17. Driscoll, J., Healy, J.: Computing fourier transforms and convolutions on the 2-sphere. *Adv. in Appl. Math.* 15, 202–250 (1994)
18. Zhou, K., Bao, H., Shi, J.: 3d surface filtering using spherical harmonics. *Computer-Aided Design* 36(4), 363–375 (2004)
19. Kostelec, P., Rockmore, D.: S2Kit: A Lite Version of SpharmonicKit, Department of Mathematics, Dartmouth College, Hanover (2004)
20. Gonzalez-Nuevo, J., Argueso, F., Lopez-Caniego, M., Toffolatti, L., Sanz, J., Vielva, P., Herranz, D.: The mexican wavelet family: Application to point source detection in cmb maps. *Mon. Not. R. Astron. Soc.* 369(4), 1603–1610 (2006)



Retraction Note to: Chapters

Abderrahim Elmoataz, Olivier Lezoray, Fathallah Nouboud,
and Driss Mammass

Retraction Note to:

Chapter “Reinstating Floyd-Steinberg: Improved Metrics for Quality Assessment of Error Diffusion Algorithms”
in: Sam Hocevar and Gary Niger: *Image and Signal Processing*,
LNCS 5099, https://doi.org/10.1007/978-3-540-69905-7_5

Retraction Note to:

Chapter “A Novel Text-Independent Speaker Verification System Using Ant Colony Optimization Algorithm”
in: Shahla Nemati, Reza Boostani, and Mohammad Davarpanah Jazi: *Image and Signal Processing*,
LNCS 5099, https://doi.org/10.1007/978-3-540-69905-7_48

For Chapter 5:

The Editors have retracted this chapter [1] because it has come to their attention that the second author, Gary Niger, is fictitious and the affiliation of the first author, Sam Hocevar, does not appear to exist. The Editors therefore no longer have confidence in the reliability of the work presented.

Sam Hocevar has not responded to correspondence regarding the concerns raised and the retraction.

[1] Hocevar, S., Niger, G.: Reinstating Floyd-Steinberg: improved metrics for quality assessment of error diffusion algorithms. In: Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D. (eds.) *Image and Signal Processing, ICISP 2008*. LNCS, vol. 5099. Springer, Berlin, Heidelberg (2008). https://doi.org/10.1007/978-3-540-69905-7_5

For chapter 48:

The paper “A Novel Text-Independent Speaker Verification System Using Ant Colony Optimization Algorithm” by Shahla Nemati, Reza Boostani and Mohammad Davarpanah Jazi, starting on page 421 of this publication, has been retracted as Sect. 3 of this paper was plagiarized from the paper starting on page 12 of LNCS 4973.

The retracted version of these chapters can be found at
https://doi.org/10.1007/978-3-540-69905-7_5
https://doi.org/10.1007/978-3-540-69905-7_48

Author Index

- Aboutajdine, Driss 321
Ahmed, Javed 128
Akhtar, Mahmood 144
Akkoul, Smail 163
Ali, Mohamed A. 463
Ambikairajah, Eliathamby 144
Amine, Aouatif 321
Amrouch, M. 447
Andersen, Hans Jørgen 85
Anguelov, Roumen 1
Armspach, Jean-Paul 67
Azzari, Pietro 413
- Bach, Francis 329
Backes, André R. 136
Bahoura, Mohammed 296
Bao, Xizhong 605
Barner, Simon 209
Beck, Dominik 281
Behrad, Alireza 378, 386
Belaroussi, Rachid 183
Benezeth, Yannick 76
Bermudez, José C.M. 527, 588
Bilane, P. 481
Bilcu, Radu Ciprian 10
Boostani, Reza 421, E1
Bouzid, Aïcha 544
Bovbel, E. 192
Brandl, Miriam 281
Bres, Stéphane 439, 481
Brun, Luc 346
Bruni, V. 254
Bruno, Odemir M. 136
Burghelea, Roxana 552
- Cabestaing, François 59
Castillo-Castaneda, Eduardo 218
Chainais, Pierre 613
Chapeau-Blondeau, François 200
Chauveau, Julien 200
Chen, Jingying 356
Choi, Tae-Sun 120
Coirault, Patrick 552
Collins, Steve 244
- Condell, J.V. 368
Constantin, Doru 312
- D’Orazio, Tiziana 263
Deschenes, François 103
Di Stefano, Luigi 413
Distante, Arcangelo 263
Donnelley, Martin 153
Douzi, Hassin 455
- Ebrahimi Atani, R. 498
Ebrahimi Atani, S. 498
Eglin, Véronique 471
El Assad, Safwan 552
El hajji, Mohamed 455
El-Hamidi, Abdallah 29
Ellouze, Noureddine 544
Elyassa, M. 447
Emile, Bruno 76, 273
Emptoz, Hubert 471, 481
Epps, Julien 144
- Faaya, Tiwuya H. 304
Fabris-Rotelli, Inger 1
Fan, Xiao 94
Ferdi, Youcef 490
- Ghofrani, S. 516
Goel, Piyush 580
- Hannuksela, Jari 405
Harba, Rachid 163, 455
Hearn, Trevor 153
Heikkilä, Janne 236, 405
Hemery, Baptiste 273
Herlin, Paulette 173
Hocevar, Sam 38
Hui, Junying 605
- Imiya, Atsushi 338
- Jafri, M. Noman 128
Jazi, Mohammad Davarpanah 421, E1
Joutel, Guillaume 471

- Kawakami Harrop Galvão, Roberto 508
 Kawulok, Michał 112
 Kheidorov, I. 192
 Khoudour, Louahdi 59
 Knoll, Alois 209
 Knowles, Greg 153
 Koenig, Emilie 613
 Kukharchik, P. 192
 Kwak, Keun-Chang 536
 Ladeev, D. 192
 Launay, Frédéric 552
 Laurent, Hélène 76, 273
 Laurini, Robert 439
 Leconge, Remy 163
 Ledee, Roger 163
 Lefèvre, Sébastien 226
 Leger, Christophe 163
 Lenz, Claus 209
 Leo, Marco 263
 Lepape, Alain 163
 Lerondel, Stéphanie 163
 Leroy, Jean-Vincent 103
 Liang, Guolong 605
 Lorenz Wendt, Frank 439
 Lugiez, Mathieu 29
 Lv, Ming 596
 Malik, Aamir Saeed 120
 Mammass, D. 447
 Mangeot, Jean Philippe 552
 Markov, Ilya 430
 Mattoccia, Stefano 413
 Mazzeo, Pier Luigi 263
 McLernnon, D.C. 516
 Ménard, Michel 29
 Messaoudi, Zahir 562
 Meurie, Cyril 59
 Miri, Sanae 67
 Mirzakuchaki, S. 498
 Montenegro M., Yasmín 527
 Morel, Guillaume 183
 Mukhopadhyay, Jayanta 580
 Müller, Thomas 209
 Murthy, K. Srikanta 46
 Nematici, Shahla 421, E1
 Nguyen, Giang Phuong 85
 Niger, Gary 38
 Noble, J. Alison 20
 Noskoski, Odair A. 588
 Ojansivu, Ville 236
 Ouldali, Abdelaziz 562
 Oussalah, Mourad 562
 Paiva, Henrique Mohallem 508
 Passat, Nicolas 67
 Pesnel, Sabrina 163
 Pham, Tuan D. 281
 Plancoulaine, Benoît 173
 Prasad, G. 368
 Pratheepan, Y. 368
 Pruvot, Jean-Hugues 346
 Rachidi, A. 447
 Rajpal, Navin 46
 Rajpoot, Kashif 20
 Rajpoot, Nasir 20
 Ratnasingam, Sivalogeswaran 244
 Revenu, Marinette 173
 Roodsarabi, Nadiya 386
 Rosenberger, Christophe 76, 273
 Rousseau, David 200
 Rziza, Mohammed 321
 Sakai, Tomoya 338
 Sangi, Pekka 405
 Shang, Xiaoxin 289
 Sheng, Xueli 605
 Signolle, Nicolas 173
 Simard, Yvan 296
 Simon, Thierry 103
 Singh, Vipula 46
 Spagnolo, Paolo 263
 State, Luminita 312
 Sun, Yan-Feng 94
 Sur, Arijit 580
 Taleb-Ahmed, Abdelmalik 490
 Tellez, Bruno 439
 Tiddeman, Bernard 356
 Tombari, Federico 413
 Torr, P.H.S 368
 Toygar, Önsen 304
 Turchioli, Christelle 218
 Turtinen, Markus 405
 Uhl, Andreas 396
 Vafadar, Maryam 378
 Vassilieva, Natalia 430

- Vehvilainen, Markku 10
Veldhuis, Raymond N.J. 289
Vert, Jean-Philippe 329
Vigneron, V. 570
Vilcahuaman, Luis 163
Vitulano, D. 254
Weber, Jonathan 226
Wild, Peter 396
- Yahiaoui, Tarek 59
Yin, Bao-Cai 94
- Zaslavskiy, Mikhail 329
Zheng, Yingxi 596
Zhong, Zhulin 596
Zhou, Bo 596
Zhou, Xiaobo 281